applied sciences

MDPI

Article

Partially versus Purely Data-Driven Approaches in SARS-CoV-2 Prediction

Samar A. Shilbayeh [1,*], Abdullah Abonamah [1] and Ahmad A. Masri [2,*]

[1] Business Analytics Program, Abu Dhabi School of Management, Abu Dhabi 0000, UAE; a.abonamah@adsm.ac.ae

[2] Medicine Department, Oregon Health & Science University, Portland, OR 97239, USA

* Correspondence: s.shilbayeh@adsm.ac.ae (S.A.S.); masria@ohsu.edu (A.A.M.); Tel.: +971-524131201 (S.A.S.)

Received: 15 July 2020; Accepted: 7 August 2020; Published: 17 August 2020

Abstract: Prediction models of coronavirus disease utilizing machine learning algorithms range from forecasting future suspect cases, predicting mortality rates, to building a pattern for country-specific pandemic end date. To predict the future suspect infection and death cases, we categorized the approaches found in the literature into: first, a purely data-driven approach, whose goal is to build a mathematical model that relates the data variables including outputs with inputs to detect general patterns. The discovered patterns can then be used to predict the future infected cases without any expert input. The second approach is partially data-driven; it uses historical data, but allows expert input such as the SIR epidemic algorithm. This approach assumes that the epidemic will end according to medical reasoning. In this paper, we compare the purely data-driven and partially-data driven approaches by applying them to data from three countries having different past pattern behavior. The countries are the US, Jordan, and Italy. It is found that those two prediction approaches yield significantly different results. Purely data-driven approach depends totally on the past behavior and does not show any decline in the number of the infected cases if the country did not experience any decline in the number of cases. On the other hand, a partially data-driven approach guarantees a timely decline of the infected curve to reach zero. Using the two approaches highlights the importance of human intervention in pandemic prediction to guide the learning process as opposed to the purely data-driven approach that predicts future cases based on the pattern detected in the data.

Keywords: purely data-driven approach; partially data-driven approach; SIR model; linear regression; exponential regression; exponential smoothing model; corona virus detection; infected cases prediction

1. Introduction

Coronavirus disease (COVID-19), which is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), emerged late in 2019 in Wuhan province in China as a series of cases with respiratory illnesses of unknown etiology at the time. Over a short period of time, the virus spread throughout the world leading to a pandemic of a magnitude that has not been seen in decades.

From the early cases in late 2019 until the writing of this manuscript (28th of June 2019), 10 million infected individuals and 500,000 deaths have been reported. More than 200 countries are affected, with major outbreaks in the United States (US), Brazil, Russia, and India [1]. The highest number of deaths are reported from the USA, Brazil, United Kingdom, Italy, and France [1]. Local policies have dictated individual countries responses, which ranged from a complete lockdown to mild restrictions in the daily activities of living. While not concrete, there seems to be a relationship between local policies and practices and the spread of the pandemic.

In view of this, forecasting the number of effected cases using machine learning approaches might inform our understanding of disease trajectory. Our goal was to compare different machine learning approaches:

(1) Purely data-drive approaches, such as linear regression [2,3] and exponential regression that correlate the dependent variable (number of cases) to one independent variable (time) [4]. Regression models concentrate in predicting the future pattern of the dependent variables using the historical pattern detected over a specific period of time. The goal of the regression models is to capture all forms of regression modeling that assumes that the data based on time series is stationary; this means that if there is no reduction in the number of infected cases in the previous data, the model forecasts that the cases will keep increasing in constant level and steady matter throughout the entire series.

(2) Partially data-driven: Susceptible-Infected-Recovered (SIR) Model, which is a type of infectious disease mathematical model, will assume that infected cases pattern will reach its peak at specific time and will decrease to reach zero case at some time in the future [5]. Using those two approaches leads to better understanding of the SARS-CoV-2 virus behavior and helps in showing the problem of dealing with uncertain situations. Considering the real-world scenarios in term of government policies, medical recommendations, and human behavior that are dynamic, it would not be wise to depend on one machine learning approach to evaluate the accuracy of the predicted model and to produce a singular answer to the question "which machine learning approach should be used in the case of uncertain, dynamic situation similar to the situation that we are currently facing in COVID-19?"

The paper is organized as follows: In Section 2, we summarize the related work. In Section 3, material and methods are explained. Section 4 covers the results, Section 5 compares the results, and Section 6 discusses the results with conclusion.

2. Related Work

Since the worldwide spread of COVID-19, researchers have engaged in intensive efforts to use scientific models and tools to predict the pandemic numbers in order to help countries understand and better manage its spread. These methods relied on variety of models, including data science, machine learning, statistical models, and SEIR models. In this section, we summarize the studies that were recently published.

The work in [6] proposed an iterative weighting for fitting Generalized Inverse Weibull distribution model to analyze and predict the growth of the COVID-19 epidemic. They applied the model using cloud computing to predict the potential threat of COVID-19 in countries worldwide in real-time. The authors in [7] examined the transmission characteristics of the COVID-19 epidemic at different stages using Gaussian distribution theory to construct a new model of coronavirus transmission. By simulating the propagation process of the COVID-19, we found that the curves of the proposed model well simulate the official data curves of Hubei, Non-Hubei area of China, and South Korea, Italy, and Iran. The authors reported key factors that affect the spread of the virus; these include the basic reproduction number, virus incubation period, and daily infection number. In [8], mathematical and numerical analysis is used to come up with reliable and accurate predictions of the COVID-19 pandemic in Pakistan. The time-dependent SIR model was used to provide future predictions. The turning point of the peak of the pandemic is defined as the day when the transmission rate becomes less than the recovering rate. They predicted that the outbreak will reach its maximum peak occurring from late May to 9th of June 2020. Their model predicted that after the peak date, the infection rate will start decreasing, but it might take months for the pandemic to completely fade away with 97% recovery happening in late August-to-September 2020. In [9], machine learning models are used to forecast the number of people affected by COVID-19. The researchers used four standard forecasting models; these are: linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) to forecast the threatening factors of COVID-19. The results show that the ES performs best among all the used models, followed by LR and LASSO, which perform well in forecasting the new confirmed cases, death rate and recovery rate, while SVM

performs poorly in all the prediction scenarios given the available dataset. In [10], a comparative analysis of machine learning and soft computing models is used to predict the COVID-19 outbreak as an alternative to SIR and SEIR models. Among a wide range of machine learning models investigated, two models showed promising results (multi-layered perceptron, MLP, and adaptive network-based fuzzy inference system, ANFIS). The study concludes that machine learning models are effective in predicting the pandemic. The study proposes to integrate machine learning and SEIR models to real novelty in the pandemic prediction.

In [11], the researchers developed a "Corona Community Tracker" using a Susceptible-Exposed-Infected-Removed (SEIR) system and used it to predict the COVID-19 pandemic in China. The authors assert that the SEIR model may help to interpret patterns of public sentiment on disseminating related health information, and assess political and economic influence of the spread of the virus. In [12], a parsimonious model that uses both quarantine of symptomatic infected individuals is proposed, as well as population-wide isolation practices in response to containment policies or behavioral changes. They show that the model explains the observed growth behavior accurately. The insights provided from the application of the model may be helpful in the implementation of containment strategies of COVID-19. In [13], epidemiological equations and data-driven neural network model is used to predict that countries in which rapid government interventions and strict public health measures for quarantine and isolation were successful in halting the spread of infection and prevent it from exploding exponentially.

In [14], the researchers proposed the SEIR model and artificial intelligence (AI) approach to predict the COVID-19 in China. They predicted that the pandemic should peak by late February, showing a gradual decline by end of April.

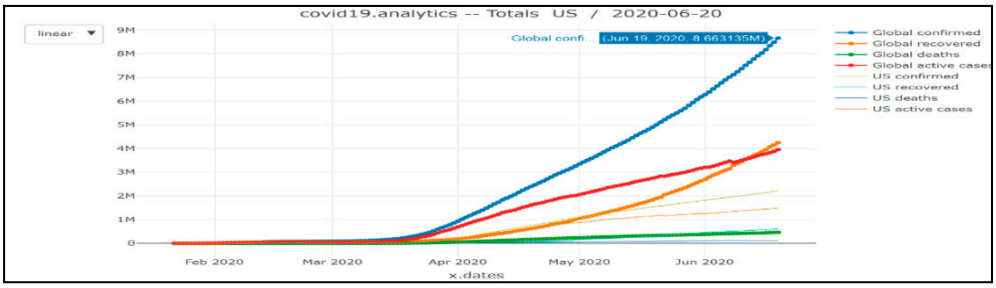
A generalized SEIR epidemiological model in [15] is proposed to predict the COVID-19 pandemic by applying a stochastic approach in fitting the model parameters using a Particle Swarm Optimization (PSO) solver. The goal of their study was to improve the reliability of predictions within 30 days. The researchers in [16] studied the use of machine learning, an artificial neural network (ANN), and a simple statistical test to identify SARS-CoV-2 positive patients from full blood counts without knowledge of symptoms or history of the individuals. The research found that with full blood counts random forest, shallow learning, and a flexible ANN model predict SARS-CoV-2 patients with high accuracy between populations on regular wards and those not admitted to hospital or in the community. Epidemiological models are used in [17] to predict the number of infected individuals and the mortality rates of the COVID-19 outbreak. However, they claimed that the available pandemic data lack essential elements and contains uncertainty, which make epidemiological models not accurate for long-term prediction. To overcome this weakness, they proposed a hybrid machine learning approach to improve the prediction. The hybrid machine learning methods combined ANFIS and multi-layered perceptron-imperialist competitive algorithm (MLP-ICA) to predict time series of infected individuals and mortality rate. The models were applied to data from Hungary and were able to predict that by late May, the outbreak and total mortality will drop substantially. In [18], an enhanced SEIR model is used to analyze the epidemic's progression of COVID-19 in eight different countries. The enhancements incorporated in the model reflect the societal feedback on pandemic and confinement features. The model with its enhancement was applied to predict the virus propagation under different conditions of confinement.

3. Material and Methods

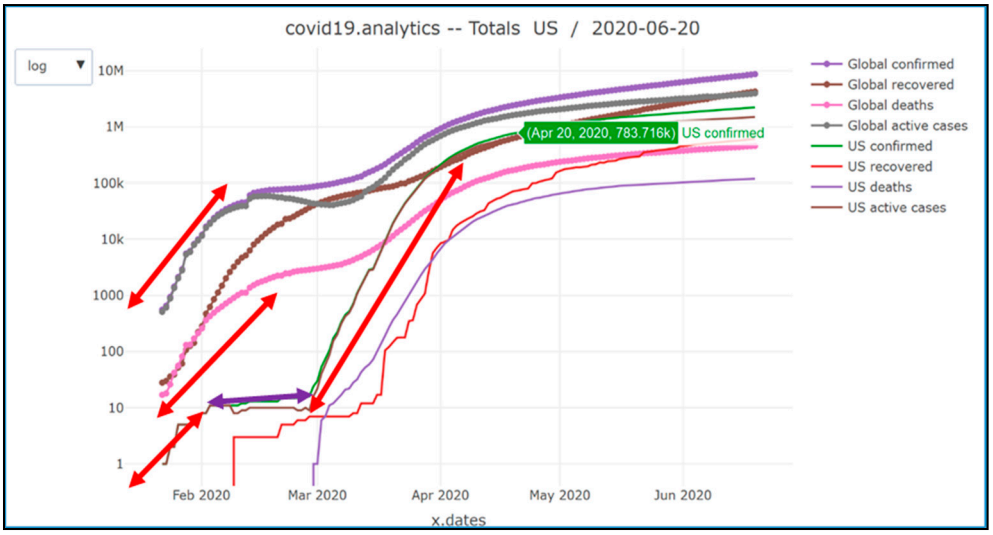
3.1. Data and Description

In this paper, we used COVID-19 data provided by John Hopkins University [1]. Specifically, our analysis focused on data from the US, Jordan, and Italy. The time period of data is from 1st of February 2020 to 26th of June 2020. The data include confirmed cases, death cases, and recovered cases; using these variables, new cases are calculated for the chosen countries.

In US, the first case of COVID-19 was reported on 22nd of January 2020 with one confirmed case; the number of cases increased to 2,220,961 cases by 19 June compared to the global confirmed cases, which were reported to be 8,663,135 cases by the same date. The number of deaths in the US was reported to be 119,112 cases on 19 June, compared to the global death cases 460,005 on the same date, which makes the US comprise of 25% of the global confirmed cases and 25.8% of the global death cases, as shown in Figure 1.



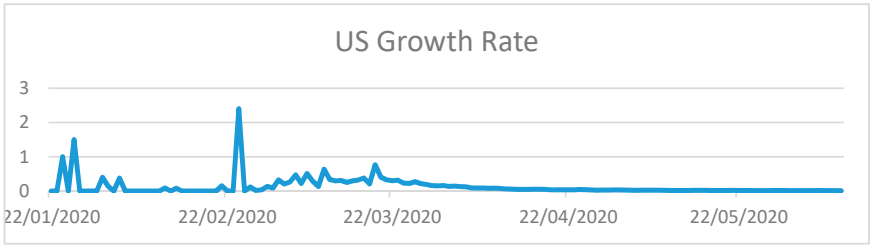
(**a**)



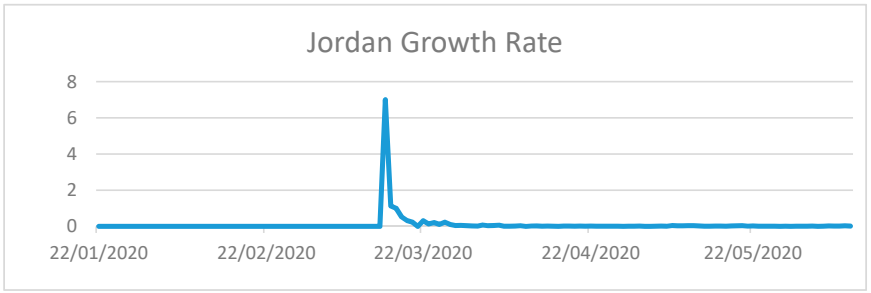
(**b**)

Figure 1. Number of confirmed/recovered/active and death cases in the US compared to the global cases: (**a**) linear scale; (**b**) logarithmic scale.

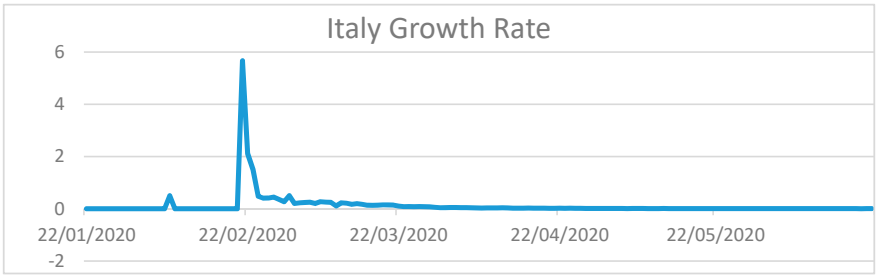
Figure 1 shows the number of confirmed, recovered, death, and active cases in US, compared to the global numbers using linear and logarithmic scales. Linear charts, which simply show the change in confirmed, deaths, active, and recovered cases over time, are not very useful in answering the question of how the speed of the outbreak compares between different countries, although it shows the increments rate for each country individually. The algorithmic scale is used to compare the exponential growth between countries that are having different starting points. As shown, parallel lines at the top and the bottom of the graph demonstrate that the US experienced the same growth rate as globally active (red double arrows lines), which obviously diminished at the end of February for the US, and on 13th of February 2020 globally (the highest US confirmed cases growth rate is 2.4% as shown in Figure 2a). On the other hand, the US confirmed cases decreased rapidly after 9th of February until 29th of February 2020 (purple arrow), and a very deep exponential increment rate is shown until 7th of April 2020; then, the curve is smoothly increased.



(**a**)



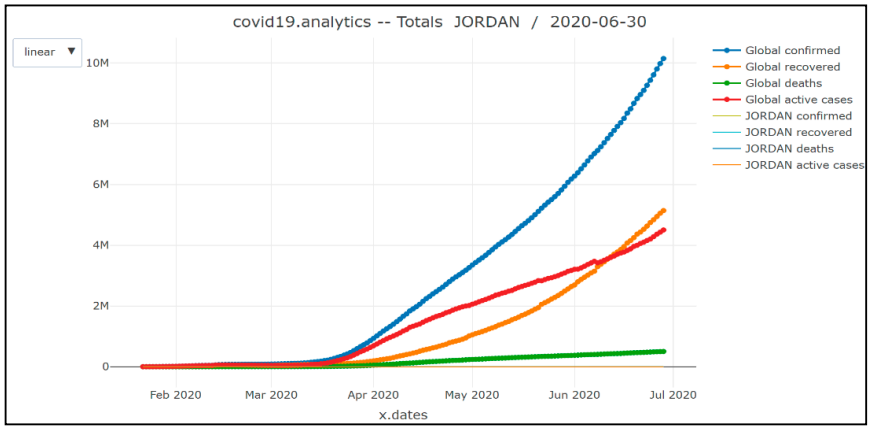
(**b**)



(**c**)

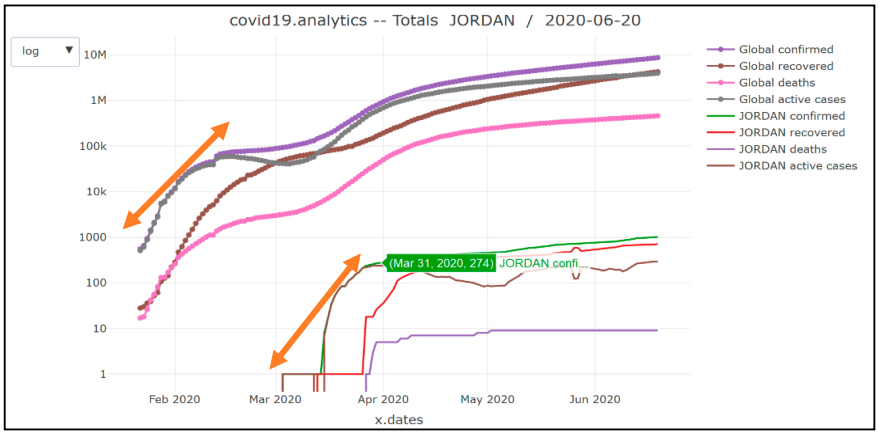
Figure 2. Confirmed cases growth rate in (**a**) US; (**b**) Jordan; (**c**) Italy.

Figure 3 shows the number of confirmed, recovered, active, and death cases for Jordan compared to the global numbers. As noted, parallel lines at the bottom of the graph show that Jordan has experienced the same confirmed cases growth rate, as globally active (orange double arrows). Jordan's deep increment ends on 15th March 2020 as illustrated in Figure 2b, which also shows that 7% is the highest confirmed cases growth rate in Jordan as of 15th March).



(**a**)

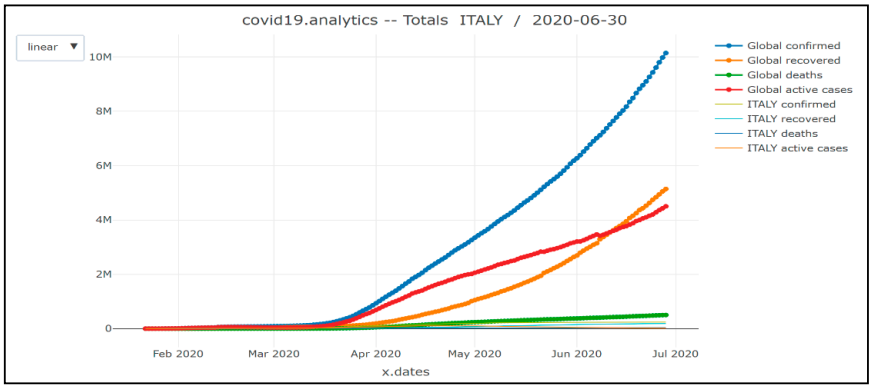
Figure 3. *Cont.*



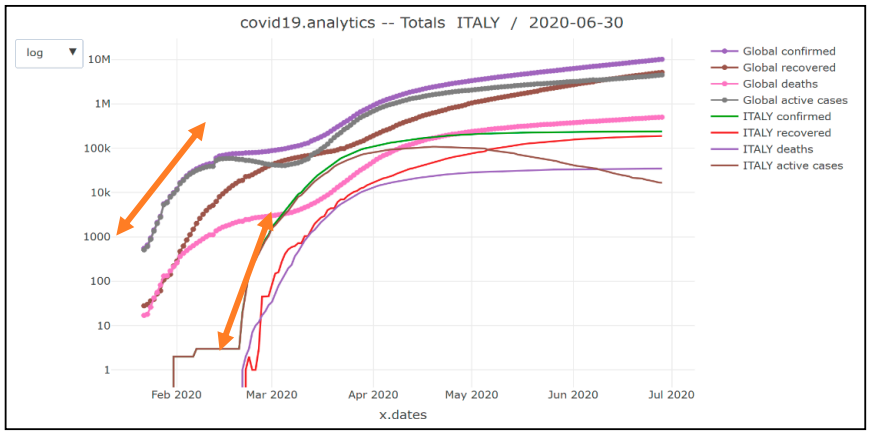
(**b**)

Figure 3. Number of confirmed/recovered/active and death cases in Jordan compared to the global cases: (**a**) linear scale; (**b**) logarithmic scale.

Figure 4 shows the number of confirmed, recovered, death, and active cases for Italy, compared to the global cases using linear and logarithmic scales. As noticed, parallel lines at the top and the bottom of the graph show that Italy has experienced a deeper exponential increment with higher slope compared to the global growth (orange double arrows lines), which contracted at the end of February for Italy (the highest confirmed cases growth rate for Italy shown in (Figure 2c) is 5%).



(**a**)



(**b**)

Figure 4. Number of confirmed/recovered/active and death cases in Italy compared to the global cases: (**a**) linear scale; (**b**) logarithmic scale.

3.2. Predictive Analytics

3.2.1. Linear Regression Model

Linear regression is a type of statistical analysis that attempts to show the relationship between two variables, dependent and independent variable. The central problem of linear regression is to infer a pattern from a set of training data and how to map new inputs of the same type to corresponding output [2]. Linear regression is widely used in prediction [19], which predicts new labels based on linear fit of the data by reducing the least square error of the predicted data points. Least squared error is measured by finding the distance between the real data points to the generated predictive model and referred to the mean square error, as shown in Figure 5. Linear regression's main aims are to first establish if there is a relation between the two sets of variables (dependent and independent). More specifically, establish if there is a statically significant relationship between them; the level of how significant the relation is can be measured using different performance measurements that will be described later in this section; second, forecast new observation for the aim of answering the following question, "can we use what we know about the relation to forecast unobserved values?"

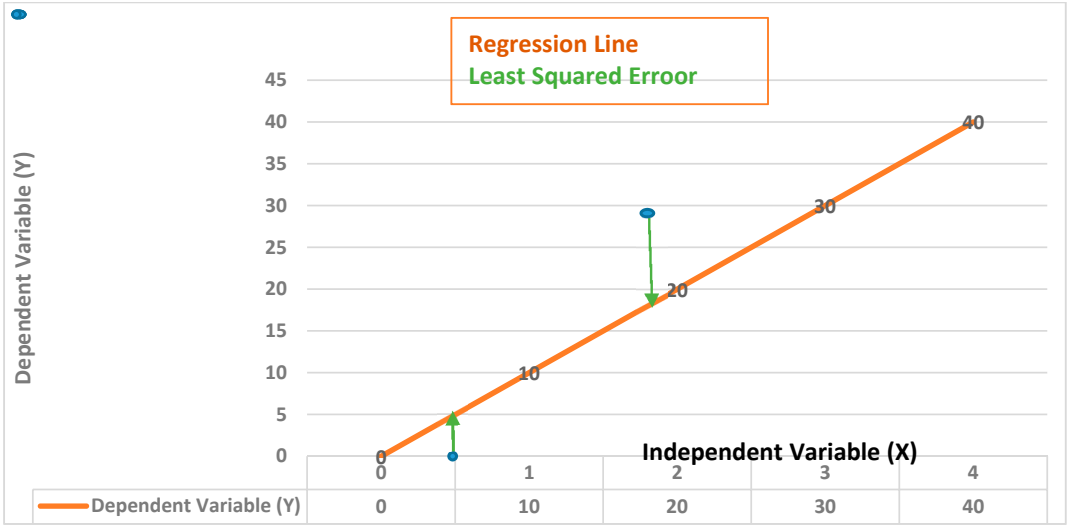


Figure 5. Linear Regression Model.

The linear regression approach may be used to predict COVID-19 through a learning model that correlates those two sets of variables (confirmed, new and death cases from one side) to the dependent variable (time on another side). The learning process can be done by finding the optimal parameters (shown in Equation (1)) for the linear fit by reducing the least squared error.

Equation (1) shows the relation between independent and dependent variable in the generated linear predictive model:

$$Y = B_0 + B_1 X + \varepsilon \quad (1)$$

where

Y: dependent variables: confirmed, new and death cases.
X: Independent variable: time
$B_0, B_1,$ and ε: Intercept, slope, and standard error respectively.
B_0: is the line intercept and defined as the average value for Y when x = 0, could be negative value.
B_1: is defined as the value increased by Y for every unit increased by X, it measures the sensitivity of Y when X increased.
ε: standered error that should be minmized.

In this paper, we used linear regression to predict the number of COVID-19new, confirmed, and death cases for a new unseen time (future time). The model confidence level is taken at 95%, so the model is significant at $p \leq 0.05$. The number of new, confirmed, and death cases are predicted for all three selected countries: US, Italy, and Jordan. The developed models are evaluated using SSE (sum of squared error), MSE (mean squared error), R Squared, and the model's significance is evaluated at p value < 0.05. The results of the linear regression models is shown and discussed in Section 4.1.

3.2.2. Exponential Regression Model

Exponential trend regression is the exponential trend line that is used to best fit the actual line when the change rate between the dependent and independent is increasing at a very high rate similar to the growth rate of the COVID-19 confirmed, new, and death cases over time. The following equations is used to express the estimated fit line [20]. The reason for using this model is as per what we mentioned in our description analysis in Section 3.1; researchers also found that the first period of the epidemic had exponential growth [21].

Equation (2) shows the relation between independent and dependent variable in the generated exponential predictive model [21]:

$$x(t) = x_0 \times b^t \tag{2}$$

- $x(t)$ is the number of cases at any given time t
- x_0 is the number of cases at the beginning, also called initial value
- b is the number of people infected by each sick person, the growth factor

We used exponential regression to predict the number of COVID-19new, confirmed, and death cases for a new unseen time (future time). With the aim of comparing this model with the linear regression model previously used, the model is evaluated at the same confiedence level, which is 95%; this model is again considered significant at $p \leq 0.05$. The number of new, confirmed, and death cases are predicted for all three selected countries—US, Italy, and Jordan. The developed model is evaluated using the same previous mentioned performance measurments, which includes SSE (sum of squared error), MSE (mean squared error), and R Squared. The results of applying the exponential regression model is shown and discussed in Section 4.2.

3.2.3. Forecasting using Exponential Smoothing Model for Time Series Data

The exponential smoothing model is a forecasting model that uses the weighted average of the past observations with the aim of forecasting new unseen values [22]. The main idea behind this model is to give more weight to the recent values in the series. Thus, the newer values will get more exponential weight than the older ones, and the weight of older data points decays exponentially to reflect their importance. In exponential smoothing methods, the error, trend, and seasonal values are combined either additively, multiplicatively, or will be left out of the model. The additive model will add the seasonal and trend result factor together, so if one of them is zero, the other will have value. In the multiplicatively model, two factors are multiplied, so if one of them is zero, then the other will be zero as a result of multiplication. This model was proposed by [23,24] and is used to forecast COVID-19 confirmed, new, and death cases.

Exponential smoothing is the most common model forecasting future realization of a time series. It requires a chosen smoothing factor with the aim of reducing the residual error to compute the average distances between the actual and predicted lines [25].

In this study, we used the exponential smoothing model to predict the number of COVID-19 new, confirmed, and death cases for a new unseen time (future time). This model suggested the data may have a trend and seasonal pattern. The forecasting has used data from 22nd of February 2020 to 20th of June 2020 to create a forecast through to 24th of August 2020, looking for a potential seasonal pattern every 7 days. The results of applying this model to the three selected countries are shown and discussed in Section 4.3.

3.2.4. SIR Epidemic Model for Spread Disease Model

The SIR (Susceptible, Infected, and Removed) model is the simplest mathematical model that can developed for infectious diseases [26]. In this model, the total population is divided into three components (shown in Figure 6) as follows:

- S: Susceptible cases—people who could potentially catch the disease; they either have low immunity system, or they still not infected.
- I: Infected—people who currently have the disease and can transmit it to others.
- R: Removed—people who have already caught the disease and have either recovered or have died.

Figure 6. SIR (Susceptible, Infected, and Removed) Model Components.

In this epidemic model, it is assumed that the epidemic is sufficiently short, so it does not last that long. In addition, the total populations remain constant for this short time period, which gives a reasonable justification for the decrement of the infected and susceptible rate, taking into consideration that all populations will get the disease. The second assumption is that the infected increment rate is propositional to the contact between susceptible and infected cases.

The SIR model can be presented using Equations (3)–(5) [26]:

$$\frac{ds}{dt} = -\frac{\beta IS}{N} \tag{3}$$

$$\frac{ds}{dt} = -\frac{\beta IS}{N} - \gamma I \tag{4}$$

$$\frac{dR}{dt} = -\gamma I \tag{5}$$

t is the time, $S(t), I(t)$, and $R(t)$ is the number of susceptible cases, infected, and recovered cases at specific time t. B is the contact rate, and $\frac{1}{\gamma}$ is the average infection rate.

We used the SIR model in this study to predict the number of COVID-19 susceptible, infected, and recovered cases for a new unseen time (future time). The prediction used data starting from 2nd of March 2020 for 250 days. The number of days considered for the initial guess is 26 in the population for each country of the study. The results of applying this model on the three selected countries are shown and discussed in Section 4.3.

4. Results

4.1. Linear Regression Model Results

A linear trend model is computed for the sum of confirmed, new, and death cases (actual and forecast) given date of day. The confidence level is taken at 95%, so the model is significant at $p \leq 0.05$; the results of applying the linear regression model for US, Jordan, and Italy are shown in Table 1.

Table 1. Linear Regression Results for Confirmed, New, and Death Cases for Italy, US, and Jordan.

Country	Italy			Jordan			US		
Model Formula	Forecast Indicator * (Day of Date + Intercept)								
Cases	Confirmed Cases	New Cases	Death Cases	Confirmed Cases	New Cases	Death Cases	Confirmed Cases	New Cases	Death Cases
SSE (sum squared error):	3.3344×10^{11}	1.21758×10^{9}	1.05497×10^{7}	1.31218×10^{6}	15846.4	12.5244	1.94177×10^{13}	2.05371×10^{10}	1.50322×10^{8}
MSE (mean squared error):	1.58781×10^{9}	5.79801×10^{6}	50,236.4	6278.39	89.0245	0.0599252	9.24653×10^{10}	9.77958×10^{10}	715,821
R-Squared:	0.846173	0.0188468	0.0227448	0.973442	0.331823	0.0072726	0.947582	0.661348	0.0280926
Standard error:	39,847.3	2407.91	224.135	79.2363	9.43528	0.244796	304,081	9889.18	846.062
p value	<0.0001	0.0458767	0.000109	<0.0001	<0.0001	0.217336	<0.0001	<0.0001	0.0145537

* The prediction is made for August 24 2020.

As shown in Table 1, SSE is the sum of the squared differences between each point and its group point's mean; it shows how variant each point is from the others. The results in Table 1 show very high variants in the three selected countries, which indicate that the confirmed, new, and death cases vary randomly in all countries, with the lowest value in Jordan death cases. MSE shows the difference between the predicted and actual values; the lowest rate is shown in Jordan death cases (i.e., 12.5). R^2 is calculated as the sum of squared errors, divided by the total sum of squared errors, as shown in Equation (6):

$$R^2 = \frac{SSM}{SST} \tag{6}$$

The larger R^2 means the ability of the predictor to predict the dependent variable. R-squared is maximum in Jordan, US, and Italy confirmed cases, respectively. This demonstrates that the time variable is a very good predictable variable for the confirmed cases being more than new and death cases. For example, 97% of the confirmed cases in Jordan can be predicted using the independent variable (time). p value < 0.05 indicates that the overall model is significant. The result shows significant models in all countries' confirmed, death, and new cases, except for the death cases model in Jordan, which means the model cannot be considered. However, the null hypothesis that assumes that there is no relation between time and death cases in Jordan is taken.

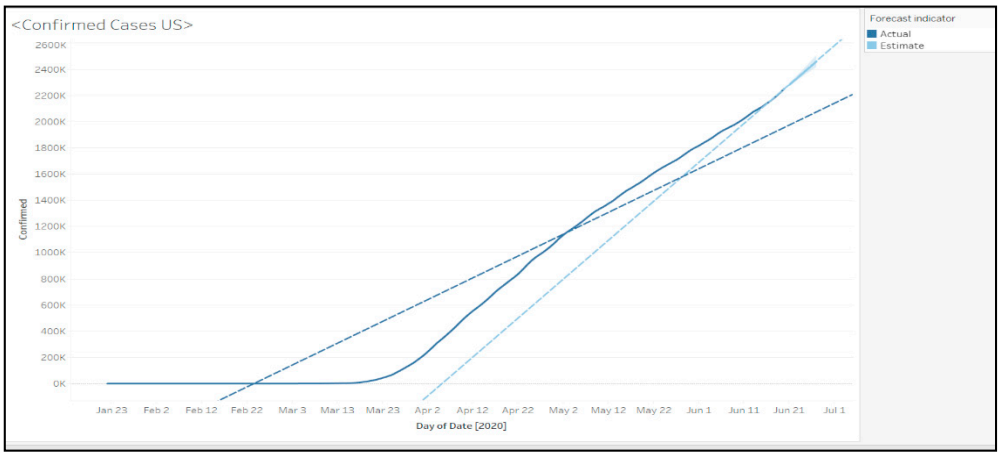
Table 2 lists the prediction numbers for confirmed, new, and death cases using the generated linear regression models. The table shows a prediction number for 24th of August 2020, and the date of first future zero cases.

Table 2. Future Prediction Numbers Using Linear Regression Models for Italy, Jordan, and US.

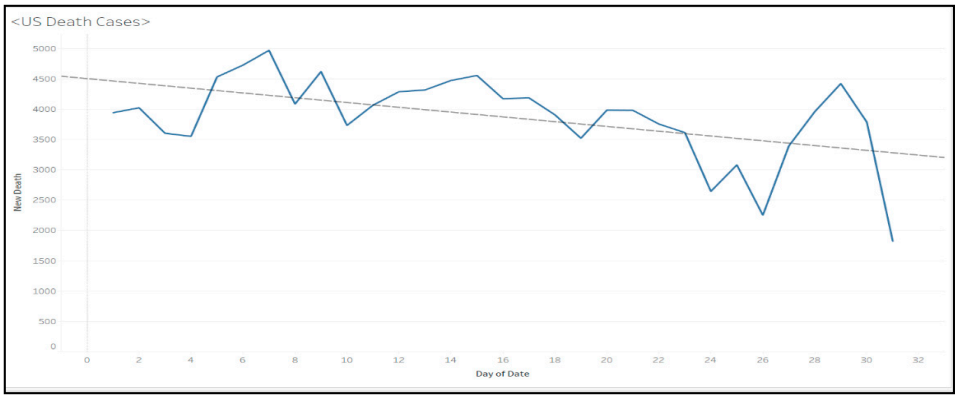
Country		Italy	Jordan	US
Confirmed Cases	Prediction (the prediction is made for 24 August 2020)	247,984	1401	4,061,952
	First zero	Not Available	Not Available	Not Available
New Cases	**Prediction**	1197	18	38,220
	First zero	17th of August 2020	Not Applicable	Not Applicable
Death Cases	**Prediction**	119	Zero	Zero
	First zero	Not Available	Started on 9th of April 2020	18th of July 2020

Table 2 shows that Italy will see zero new cases from 17th of August 2020; zero death cases in Jordan began from 9th of April 2020 and will be seen in the US from 18th of July 2020. The results show no zero cases for confirmed cases in the future for the selected countries using the linear regression model.

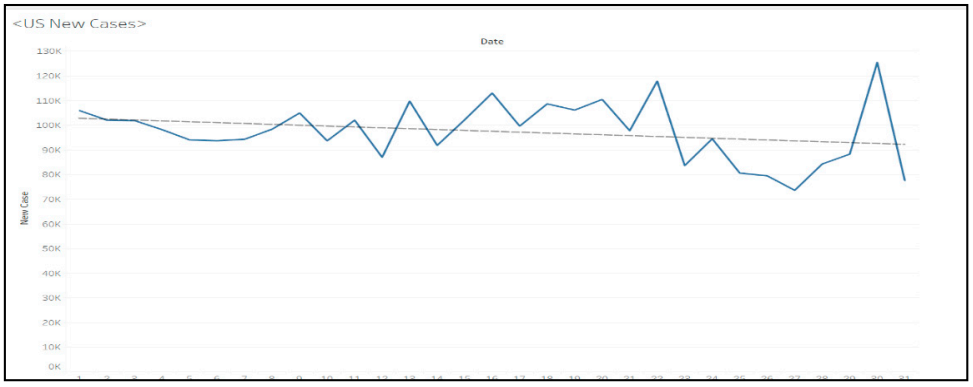
Figures 7–9 show the linear regression models for all cases in the three selected countries.



(**a**)

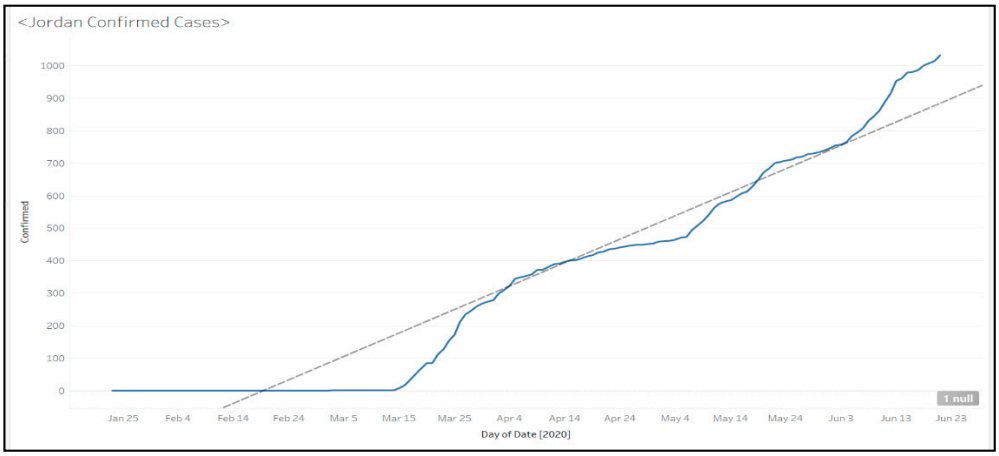


(**b**)

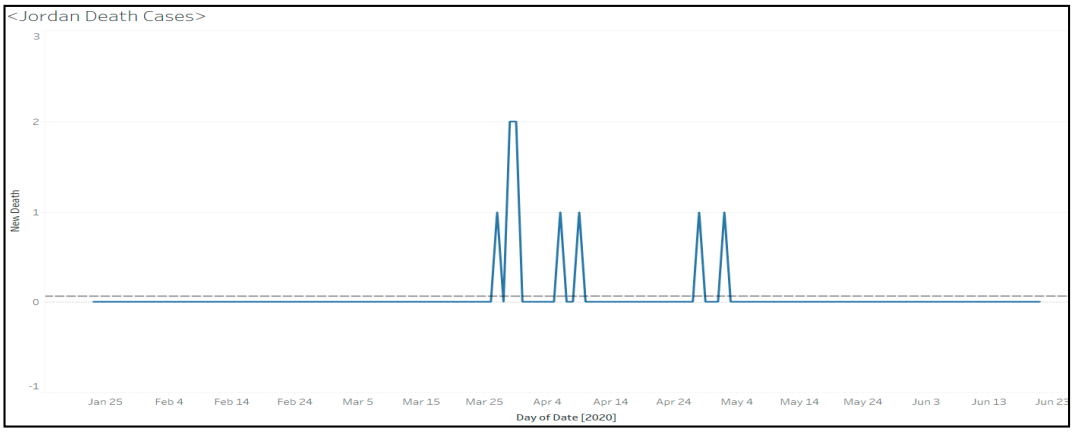


(**c**)

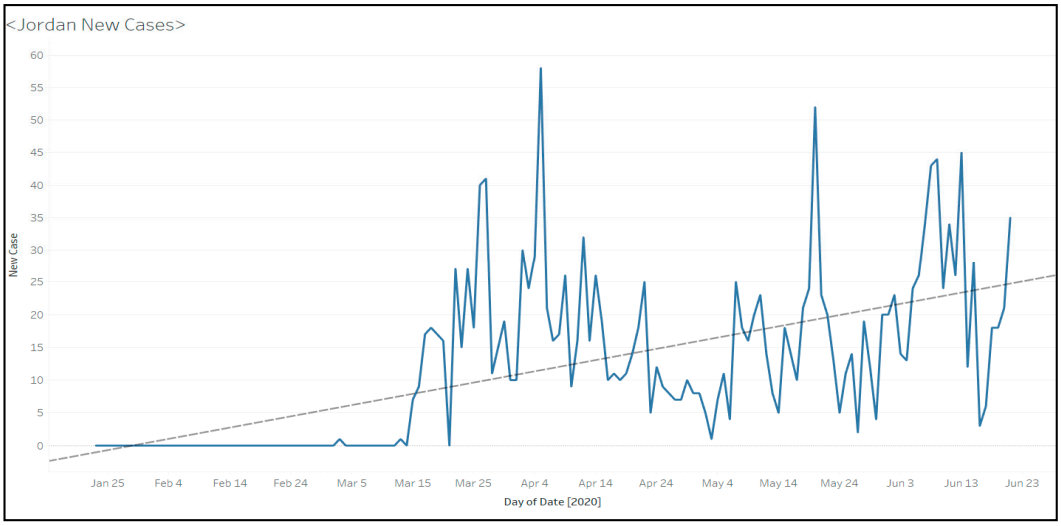
Figure 7. Linear regression model results for (**a**) confirmed cases; (**b**) death cases, and (**c**) new cases in the US.



(**a**)

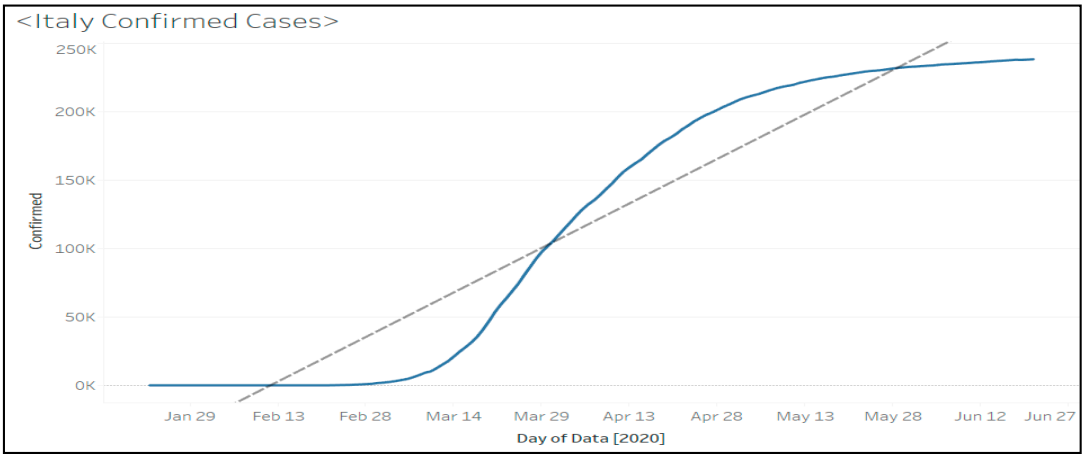


(**b**)

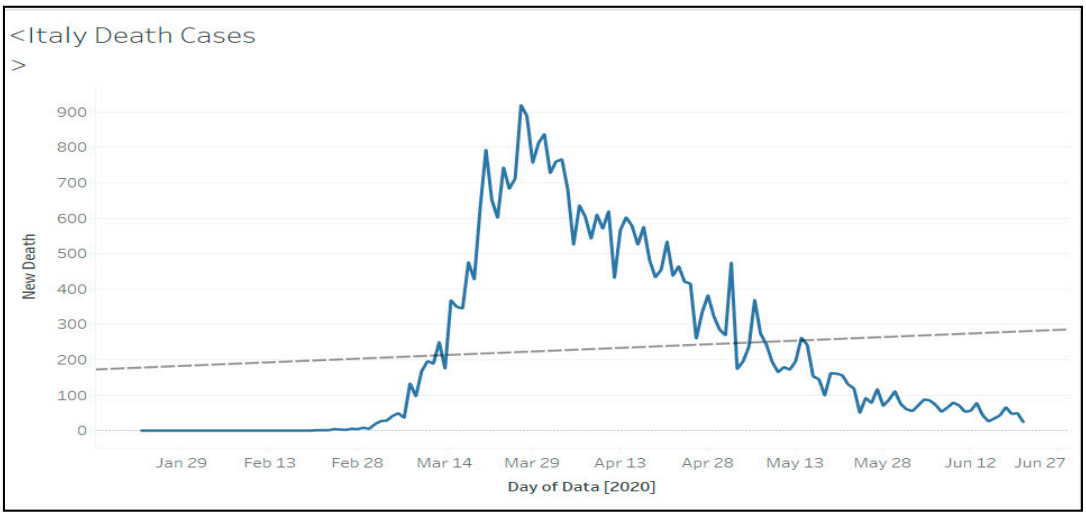


(**c**)

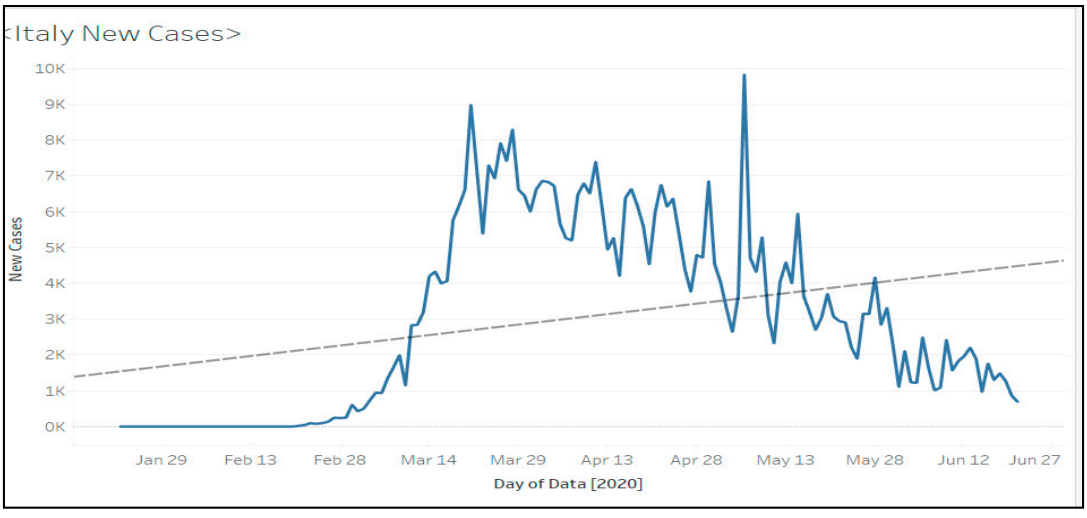
Figure 8. Linear regression model results for (**a**) confirmed cases, (**b**) death cases; and (**c**) new cases in Jordan.



(**a**)



(**b**)



(**c**)

Figure 9. Linear regression model results for (**a**) confirmed cases, (**b**) death cases, and (**c**) new cases in Italy.

4.2. Exponential Regression Model Results

In this model, all dependent variables are transformed into log scale before applying the regression model. An exponential trend model is computed for the sum of confirmed, new, and death cases (actual and forecast) given day of date. The confidence level is taken at 95%, so the model is significant at $p \leq 0.05$; the results of applying the exponential regression model for US, Jordan, and Italy are shown in Table 3.

Table 3. Exponential Regression Results for Confirmed, New, and Death Cases in Italy, US, and Jordan.

Country	Italy			Jordan			US		
Model Formula	Forecast Indicator Day of Date + intercept)								
Cases	Confirmed Cases	New Cases	Death Cases	Confirmed Cases	New Cases	Death Cases	Confirmed Cases	New Cases	Death Cases
SSE (sum squared error):	1240.22	311.656	308.051	265.67	59.9513	0.498341	1279.86	826.514	484.552
MSE (mean squared error):	6.17027	1.80148	1.71139	1.57201	0.379438	0.0996681	6.09456	4.49193	3.64325
R-Squared:	0.534905	0.0042252	0.0093464	0.537175	0.0507974	0.273939	0.751734	0.47227	0.114924
Standard error:	2.484	1.34219	1.3082	1.2538	0.615986	0.315703	2.46872	2.11942	1.90873
p value	<0.0001	0.392753	0.194185	<0.0001	0.0041635	0.227987	<0.0001	<0.0001	<0.0001

As shown in Table 3, the lowest SSE value is recorded in Jordan's death cases, which is very similar to the result of the linear regression model; this shows that the death cases in Jordan are varying at a very low random rate. Figure 9 shows the constant exponential line for death cases in Jordan. MSE shows the different between the predicted and actual values. The exponential regression model outperforms the linear regression model, as exhibited by the low MSE rates recorded in all confirmed, new, and death cases in the three selected countries.

R-squared is maximum in the US, Jordan, and Italy confirmed cases, respectively. This shows that the time variable is a very good predictable variable for the confirmed cases more than new and death cases. For example, 75% of the confirmed cases in the US can be predicted using the independent variable (time).

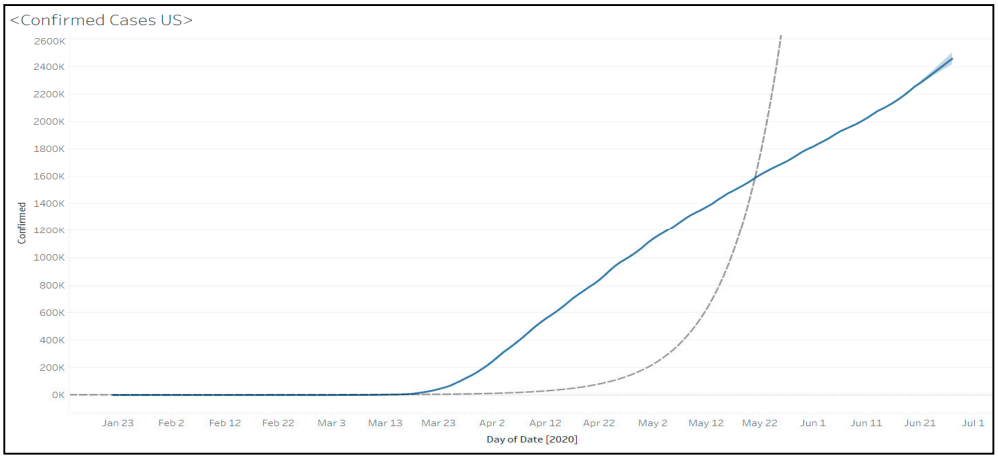
p value < 0.05 indicates that the overall model is significant. The result shows significant models in all countries confirmed, death, and new cases, except the death cases model in Jordan, which means the model cannot be taken. However, the null hypothesis that assumes that there is no relation between time and death cases in Jordan is taken.

Table 4 lists the prediction numbers for confirmed, new, and death cases using the generated exponential regression models. The table shows the prediction number for 24 August 2020, and the date of first future zero cases.

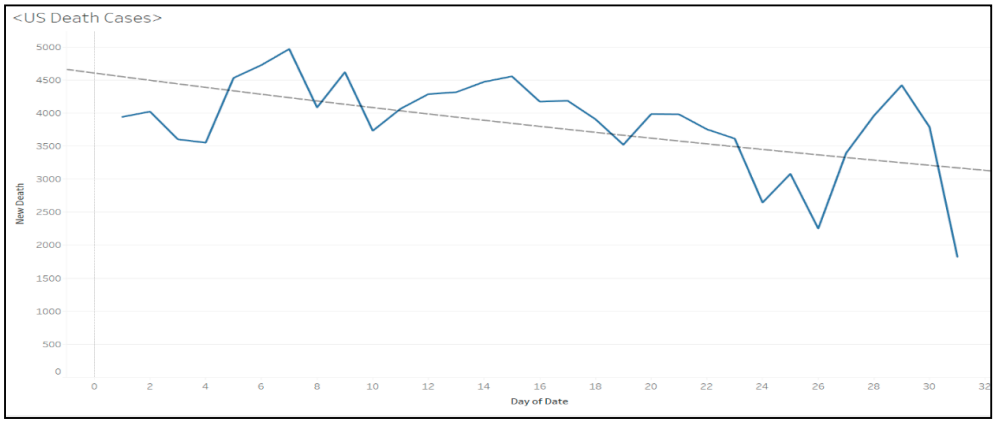
Table 4. Future Prediction Numbers using Exponential Regression Models for Italy, Jordan, and US.

Country		Italy	Jordan	US
Confirmed Cases	**Prediction (done on 24 August 2020)**	247,984	1401	4,061,952
	First zero	Not Applicable	Not Applicable	Not Applicable
New Cases	**Prediction**	1197	18	38,221
	First Zero	26 July 2020	Not Applicable	Not Applicable
Death cases	**Prediction**	119	Zero	Zero
	First Zero	Not Applicable	Zero Since 2nd May 2020	12th July 2020

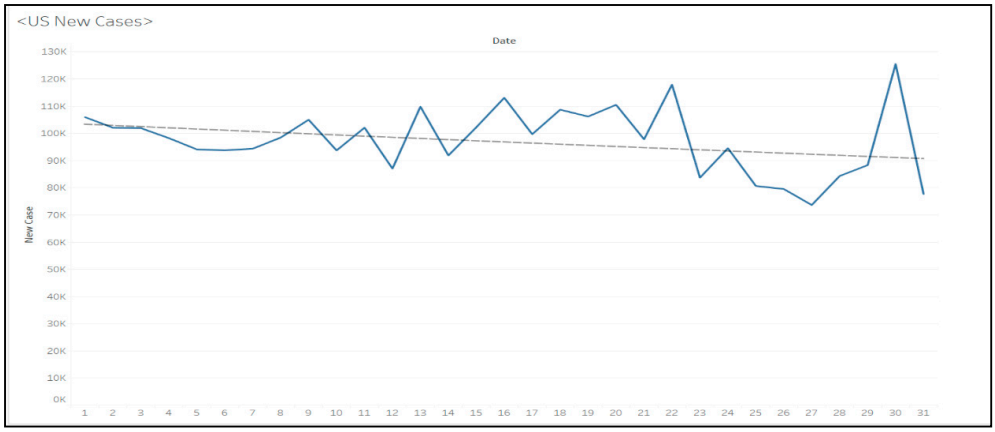
Table 4 shows that Italy will see zero new cases from 26th July 2020; zero death cases began in Jordan from 2nd, May 2020 and will be seen in the US from 12th July 2020. The results show no zero cases for confirmed cases in the future for the selected countries using the exponential regression model. Figures 10–12 show the exponential regression models for all cases in the three selected countries.



(**a**)

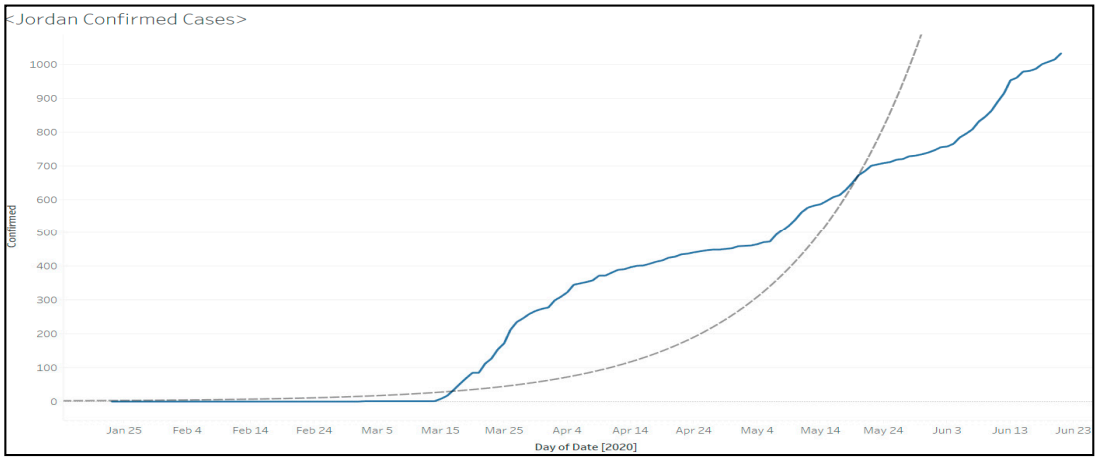


(**b**)

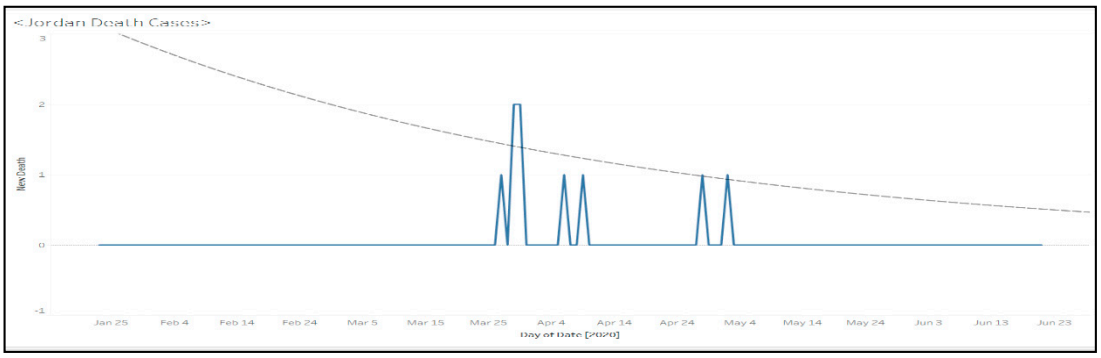


(**c**)

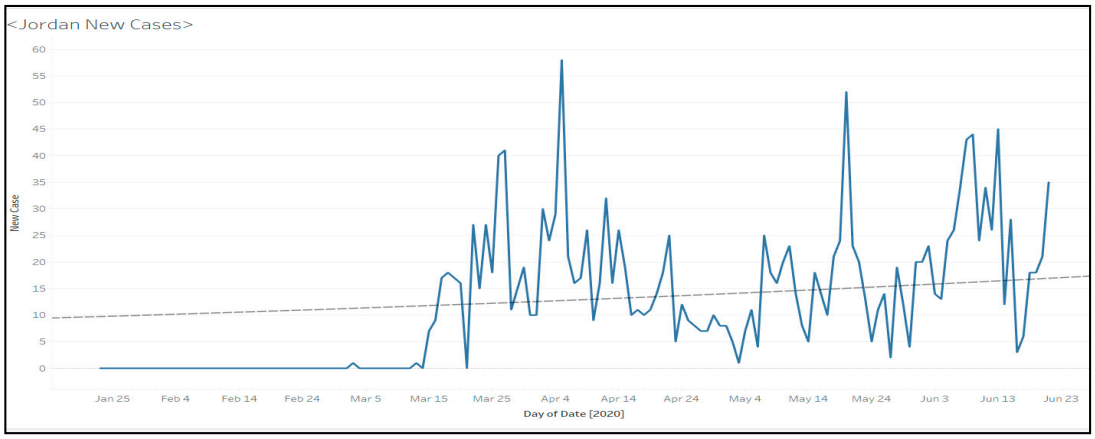
Figure 10. Exponential Regression Models for (**a**) Confirmed Cases; (**b**) Death Cases; (**c**) New Cases in US.



(**a**)

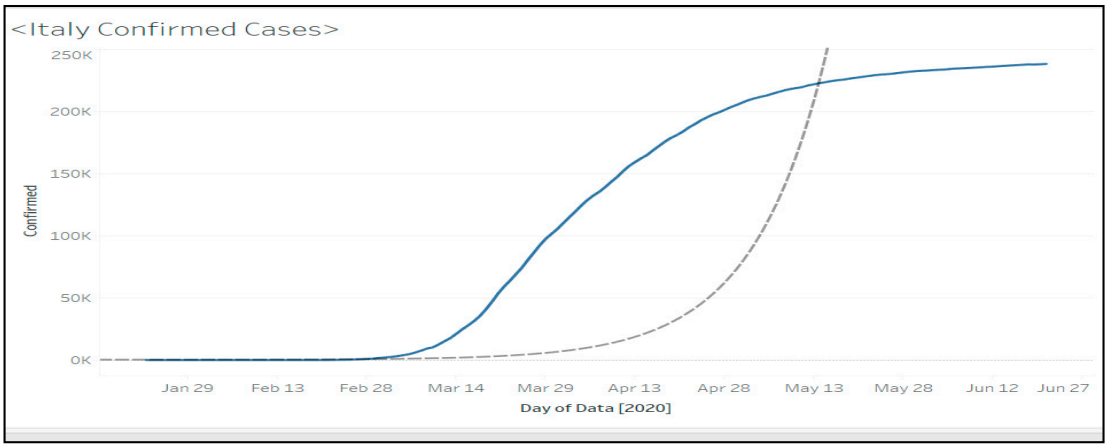


(**b**)

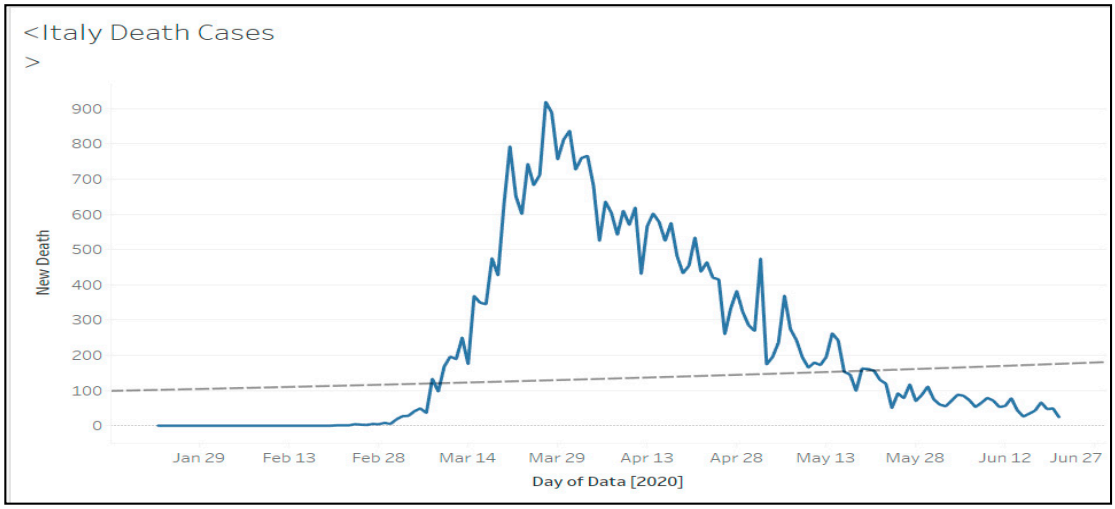


(**c**)

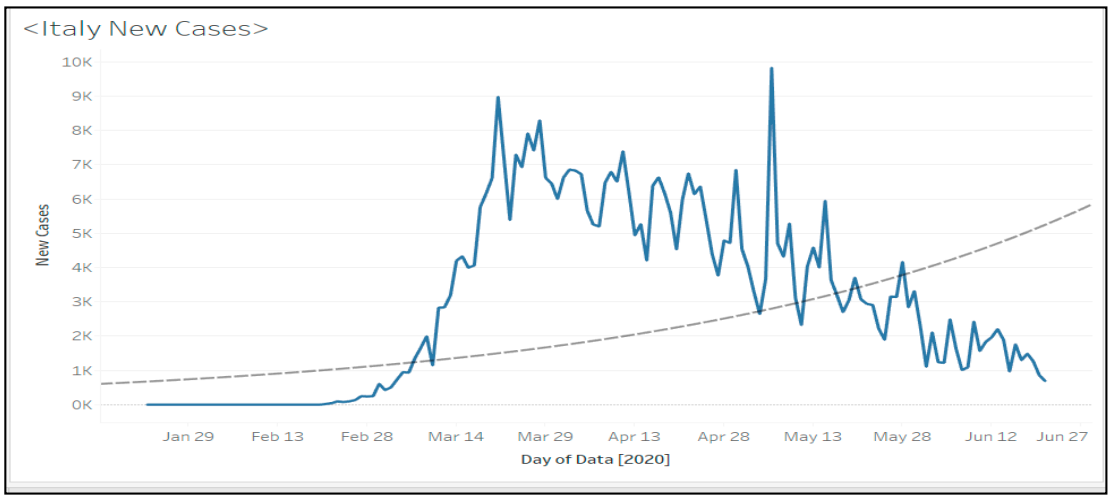
Figure 11. Exponential Regression Models for (**a**) Confirmed Cases; (**b**) Death Cases; (**c**) and New Cases in Jordan.



(**a**)



(**b**)



(**c**)

Figure 12. Exponential Regression Models for (**a**) Confirmed Cases; (**b**) Death Cases; and (**c**) New Cases in Italy.

4.3. Forecasting Using Exponential Regression Model

The results of applying the exponential smoothing model for forecasting COVID-19 confirmed, new and death cases between the period of 22nd of February to 20th June 2020 is shown in Tables 5–7. The forecasting is done for 65 days (from 21st June to 24th of August 2020). The forecasting assumed that the COVID-19 data has a trend and/or seasonal pattern using the addictive model previously explained in Section 3.2.3.

Table 5. Forecasting Results for Confirmed Cases using the Exponential Smoothing Model.

Forecast Forward	65 Days (21st June 2020–24th of August 2020)								
Forecast Based on	22nd of February 2020–20th of June 2020								
	Initial			Change from Initial	Seasonal Effect		Contribution		
Country	21st of June 2020			21st of June 2020–24th of August 2020	High	Low	Trend	Season	Quality
Jordan	1022	±	1.8%	37.1%	100.0%		100.0%	0.0%	OK
Italy	238,505	±	0.7%	4.2%	100.0%		100.0%	0.0%	Good
US	2,279,626	±	0.5%	78.2%	100.0%		100.0%	0.0%	Good

Table 6. Forecasting Results for New Cases using the Exponential Smoothing Model.

Forecast Forward	65 Days (21st of June 2020–24th of August 2020)									
Forecast Based on	22nd February 2020–20th of June 2020									
	Initial			Change from Initial	Seasonal Effect			Contribution		
Country	21st of June 2020			21st of June 2020–24th of August 2020	High		Low	Trend	Season	Quality
Jordan	18	±	103.8%	0.0%	None			0.0%	0.0%	Poor
Italy	395	±	490.7%	202.6%	14th August 2020	57.2%	17th of August 2020	−116.1%	0.0% 10.0%	Ok
US	38,221	±	39.1%	0.0%	None		0.0%		0.0%	Poor

Table 7. Forecasting Results for Death Cases using the Exponential Smoothing Model.

Forecast Forward	65 days (21st of June 2020–24th of August 2020)									
Forecast Based on	22nd of February 2020–20th of June 2020									
	Initial			Change from Initial	Seasonal Effect			Contribution		
Country	21st of June 2020			21st of June 2020–24th of August 2020	High		Low	Trend	Season	Quality
Jordan	0	±	156,004.5%	0.0%	None		None	0.0%	0.0%	Poor
Italy	45	±	293.7%	173.6%	None		None	100%	0.0%	Poor
US	170	±	266.9%	−541.6%	August 19	−16.8%	16th of August 2020	49.1%	80.2% 19.8%	Ok

Applying the exponential smoothing model on the three selected countries to forecast the number of confirmed, new, and death cases results in developing good forecasting models for confirmed cases with 100% trend and no seasonal effects for all confirmed cases, as shown in Table 5. The new cases forecasting results in the development of an acceptable forecasting model for Italy and poor models for US and Jordan. On the other hand, the US death cases forecasting results in an acceptable death cases prediction model with 80% trend and 20% seasonal effect. The highest seasonal effect for death cases in

US is shown on August 19, and the lowest is shown in 16th of August 2020, as shown in Table 8. The following is the forecasting numbers for all cases using the exponential smoothing model. Forecasting is done only if the model generated is good or okay (highlighted cells), and neglected if a poor model is generated.

Table 8. Future Forecasting using the Exponential Smoothing Model for Italy, Jordan, and US.

Country		Italy	Jordan	US
Confirmed Cases	**Prediction**	248,616	1427	4,180,774
	First zero	Not Applicable	Not Applicable	Not Applicable
New Cases	**Prediction**	Zero	Poor Model	Poor Model
	First Zero	6th of July 2020	Poor Model	Poor Model
Death Cases	**Prediction**	Poor Model	Poor Model	Zero
	First Zero	Poor Model	Poor Model	5th of July 2020

Table 8 shows that Italy will see zero new cases from 6th of July 2020 and the US will have zero death cases from 5th of July 2020. The results show no zero case for confirmed cases in the future for the selected countries using the exponential smoothing models for the selected period.

4.4. SIR Epidemic Model Results

Following is the results of applying the SIR model to forecast COVID-19 susceptible/infected and recovered cases for US, Italy, and Jordan for 400 days starting from 1st of February 2020.

The results show decrements in susceptible and infected cases (blue and red curves) in Figures 11 and 12, and increment in recovery cases (green curves) in the same diagrams.

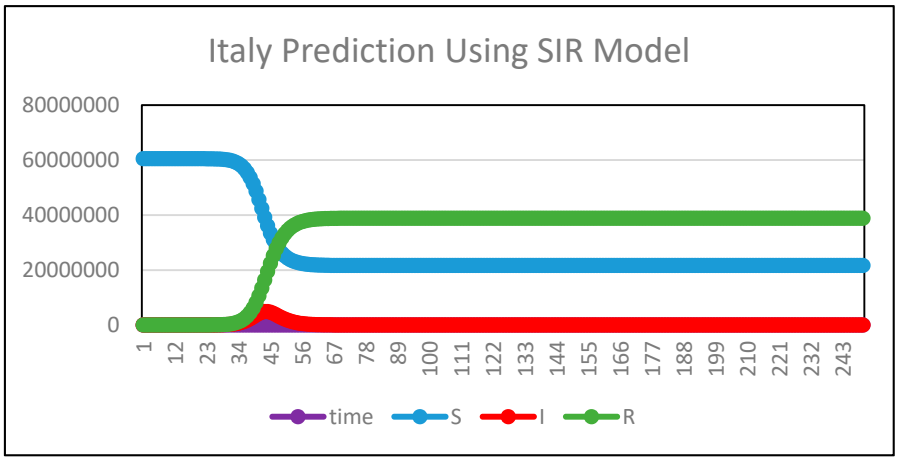
The number of predicted cases using the SIR Model with date zero cases is shown in Table 9.

Table 9. Future Forecasting using the SIR Model for Italy, Jordan, and US.

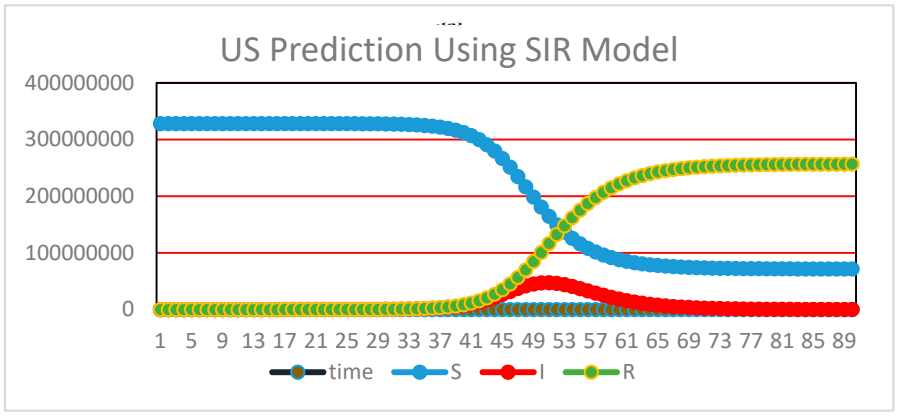
Country		Italy	Jordan	US
Susceptible	**Prediction (done on 24th of August 2020)**	21,629,365	5,701,847	71,555,047
	First zero	Not Applicable	Not Applicable	Not Applicable
Infected	**Prediction**	Zero	20	Zero
	First Zero	10th of June 2020	20th of August 2020	1st of August 2020
Recovered	**Prediction**	38832460	4501872	256645000
	First Zero	21,629,365	5,701,847	71,555,047

The table shows zero infected cases in Italy on 10th of June 2020, in Jordan on 20th of August 2020, and in the US on August 1. It also shows the predicted numbers for susceptible and recovered cases on 24th of August 2020.

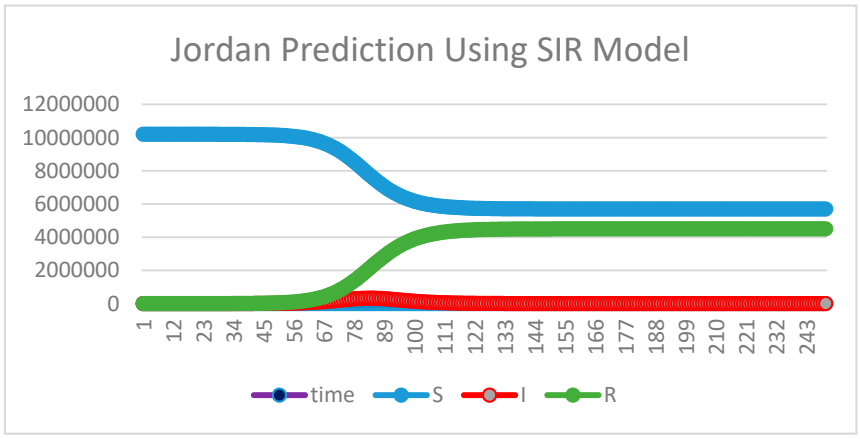
Table 9 shows that zero infected cases in Italy on 10th of June 2020, in Jordan on 20th of August 2020, and in the US on 1st of August 2020 (Figure 9a–c, red curves). The results show that susceptible cases will decrease in all countries, as shown in Figure 9a–c, (blue curves), but will not reach zero. However, the recovered cases will increase in all countries, as shown in Figure 13a–c (green curves).



(**b**)



(**c**)



(**d**)

Figure 13. SIR Models Prediction for Susceptible, Recovered, and Infected Cases: (**a**) Italy; (**b**) US; and (**c**) Jordan.

5. Comparison of the Results

This section compares the results of forecasting of the numbers of future confirmed, new, and death cases in Italy, Jordan, and US using linear, exponential regression, and exponential smoothing models. These three "purely data driven models" are used to predict the number of future cases on 24th of August 2020. The date of future zero cases are also forecasted using the three mentioned models. The results are shown in Table 10.

Table 10. Future Forecasting Using Linear, Exponential, and Exponential Smoothing Models in Italy, Jordan, and US.

		Linear Regression			Exponential Regression			Exponential Smoothing		
		Italy	**Jordan**	**US**	**Italy**	**Jordan**	**US**	**Italy**	**Jordan**	**US**
Confirmed Cases	Prediction (The prediction is done on 24th of August 2020	247,984	1401	4,061,952	247,984	1401	4,61,952	248,616	1427	4,180,774
	First zero	NA	NA	NA	NA	NA	NA	NA	NA	NA
New Cases	Prediction	1,197	18	38,220	1197	18	38,221	Zero	Poor Model	Poor Model
	First Zero	17th of August 2020	NA	NA	26th of July 2020	NA	NA	6 the of July 2020	Poor Model	Poor Model
Death Cases	Prediction	119	Zero	Zero	119	Zero	Zero	Poor Model	Poor Model	Zero
	First Zero	NA	9th of April 2020	18th of July 2020	NA	2nd of May 2020	12th of July 2020	Poor Model	Poor Model	5th of July 2020

As shown in Table 10, no first zero confirmed case was detected using the generated models in the three selected countries, while the first zero new case was forecasted in Italy to be on 17th of August 2020 using the linear regression model, and had started on 26th of July 2020 using the exponential regression model, and on 6th of July 2020 using the exponential smoothing model. On the other hand, the first zero death case was detected in Jordan on 9th of April 2020 and in the US on 18th of July 2020 using the linear regression model. Using the exponential regression model, the first zero death case was detected in Jordan and the US on 2nd of May 2020 and 12th of July 2020, respectively. The first zero death case was detected in the US on 5th of July 2020 using the exponential smoothing model.

Figure 14 shows the number of future confirmed cases predicted for Italy, Jordan, and US using the three selected models. It is worth mentioning here that a zero confirmed case was not found in future prediction using the purely data-driven models, in contrast with the SIR model, where a zero infected case was found on 10th of June 2020 for Italy, and predicted to be on 1st of August 2020 for the US and 20th of August 2020 for Jordan respectively, as listed in Table 9 and shown in Figure 14.

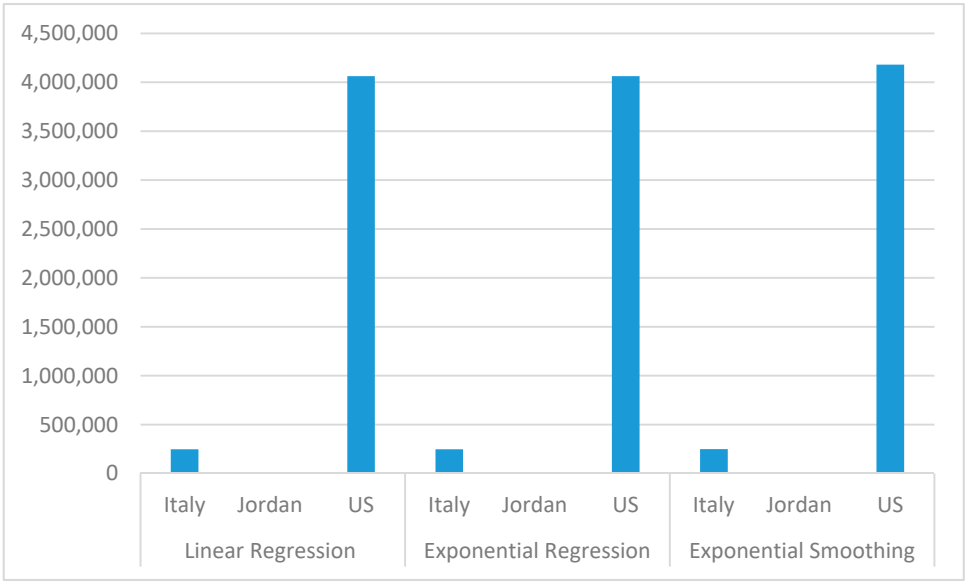


Figure 14. Confirmed case prediction for 24th of August 2020 using the Linear, Exponential Regression, and Exponential Smoothing models in Italy, Jordan, and US.

Figure 15 shows zero new cases forecasted in Italy on 17th of August 2020 using the linear regression model, on 26th of July 2020 using the exponential regression model, and 6th of July 2020 using the exponential smoothing model. Zero new cases are not forecasted in Jordan or the US using the three purely data-driven models.

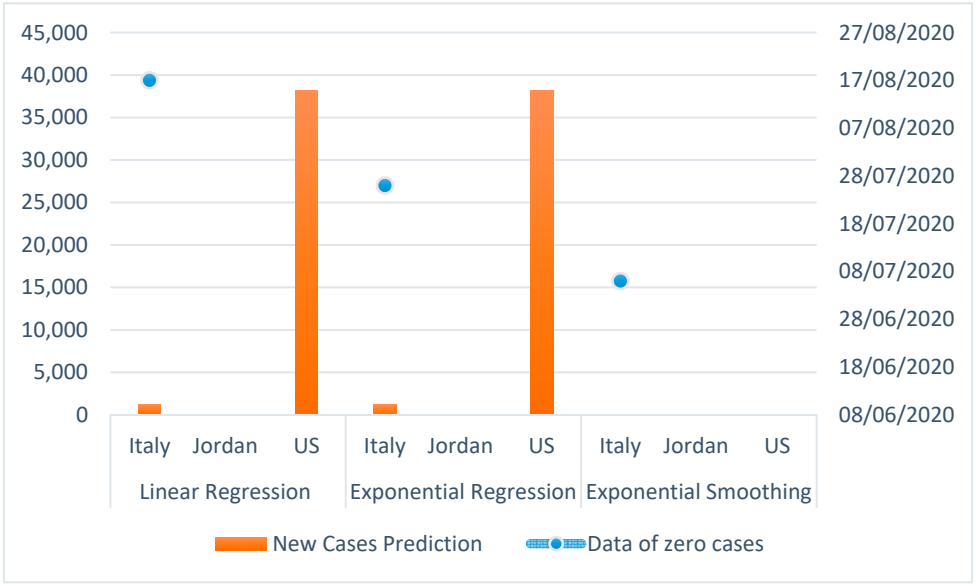


Figure 15. New case prediction for 24th of August 2020 and zero new case prediction using the linear, exponential regression, and exponential smoothing models in Italy, Jordan, and the US.

Figure 16 shows zero death cases detected Jordan and the US on 9th of April and 18th of July 2020 using the linear regression model; on 2nd of May and 12th of July 2020 using the exponential regression model; and on 5th of July 2020 using the exponential smoothing model, respectively. Zero death cases are not forecasted in Italy at any time using the three selected models.

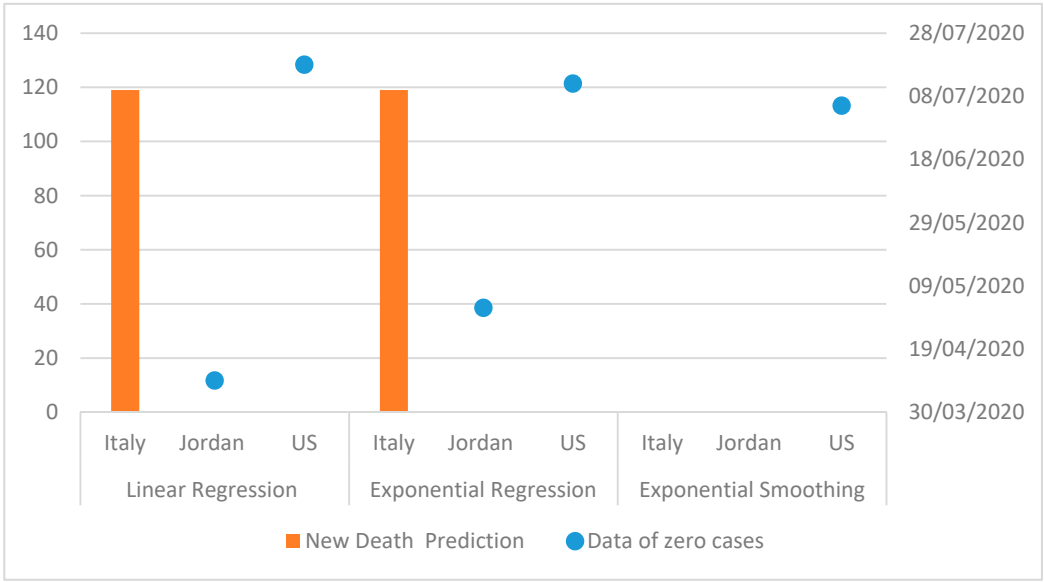


Figure 16. Death case prediction for 24th of August 2020 and zero case prediction using the linear, exponential regression, and exponential smoothing models in Italy, Jordan, and the US.

6. Discussion and Conclusions

Our study highlights the advantages and disadvantages of different modeling strategies related to the COVID-19 pandemic. With or the aim of forecasting the future number of confirmed, new, and death cases in three selected countries, two machine learning approaches were utilized and compared. The first approach, which is purely data-driven, aims at capturing previous data behavior and assumes that future patterns are just a repetition of the past without considering any human intervention. This approach failed to find any future zero cases for countries that did not experience any decline in their previous COVID-19 number curves. The second approach, which we refer to as partially data-driven, aims to build a model that can predict future cases by allowing some human intervention. This approach assumes that infected cases will reach their peak at specific time and will decrease to reach zero cases at a specific point. Utilizing this approach shows that zero cases are reached even for the countries that did not witness any curve reduction. Linear, exponential, and forecasting using exponential smoothing models are used for the first approach, and the SIR model is used from the second one.

In the machine learning community, there is a general agreement that there is no one machine learning approach that will work best all the time. Taking into consideration uncertain situations like pandemics and with multivariate factors such as human restrictions, government policies, expert recommendations, and human behavior that are dynamically changeable, we tried to answer the question: "Which machine learning approach will best predict the growth and trend of the COVID-19 pandemic?"

The concepts addressed here are still a work in progress, influenced by the rapid changes that are happening in the epidemiology and trajectory of COVID-19. Our study provides important comparative analytical and modeling approaches to the current pandemic, which would help push the envelope towards narrowing the choice for best modeling strategies for COVID-19.

Author Contributions: Conceptualization, S.A.S. and A.A.; methodology, S.A.S.; software, S.A.S.; validation, S.A.S.; formal analysis, S.A.S.; investigation, S.A.S. and A.A.M.; resources, S.A.S. and A.A.; data curation, S.A.S.; writing—original draft preparation, S.A.S.; writing—review and editing, A.A. and A.A.M.; visualization, S.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: Ahmad Masri received research grants from Pfizer and Akcea (paid to OHSU) and serves on an advisory board with Ionis. None of these have conflicts related to this work, and the other authors have no conflict of interest.

References

1. WHO Coronavirus Disease (COVID-19) Dashboard. Available online: https://covid19.who.int/?gclid=EAIaIQobChMI-5Coica_6gIVhIbVCh1lpA4DEAAYASAAEgJdHvD_BwE (accessed on 8 July 2020).
2. Seber, G.A.; Lee, A.J. *Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012; Volume 329.
3. Ghosal, S.; Sengupta, S.; Majumder, M.; Sinha, B. Prediction of the number of deaths in India due to SARS-CoV-2 at 5–6 weeks. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2020**. [CrossRef] [PubMed]
4. Pasayat, A.K.; Pati, S.N.; Maharana, A. Predicting the COVID-19 positive cases in India with concern to Lockdown by using Mathematical and Machine Learning based Models. *medRxiv* **2020**. [CrossRef]
5. Zhu, K.; Ying, L. Information Source Detection in the SIR Model: A Sample-Path-Based Approach. *IEEE/ACM Trans. Netw.* **2014**, *24*, 408–421. [CrossRef]
6. Tuli, S.; Tuli, S.; Tuli, R.; Gill, S.S. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet Things* **2020**, *11*, 100222. [CrossRef]
7. Li, L.; Yang, Z.; Dang, Z.; Meng, C.; Huang, J.; Meng, H.T.; Wang, D.; Chen, G.; Zhang, J.; Peng, H.; et al. Propagation analysis and prediction of the COVID-19. *Infect. Dis. Model.* **2020**, *5*, 282–292. [CrossRef]
8. Waqas, M.; Farooq, M.; Ahmad, R.; Ahmad, A. Analysis and Prediction of COVID-19 Pandemic in Pakistan using Time-dependent SIR Model. *arXiv* **2020**, arXiv:2005.02353.

9. Rustam, F.; Reshi, A.A.; Mehmood, A.; Ullah, S.; On, B.-W.; Aslam, W.; Choi, G.S. COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access* **2020**, *8*, 101489–101499. [CrossRef]
10. Ardabili, S.F.; Mosavi, A.; Ghamisi, P.; Ferdinand, F.; Varkonyi-Koczy, A.R.; Reuter, U.; Rabczuk, T.; Atkinson, P.M. COVID-19 Outbreak Prediction with Machine Learning. *SSRN Electron. J.* **2020**. [CrossRef]
11. Hamzah, F.A.B.; Lau, C.H.; Nazri, H.; Ligot, D.V.; Lee, G.; Tan, C.L.; Shaib, M.K.B.M.; Zaidon, U.H.B.; Abdullah, A.B.; Chung, M.H.; et al. CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction. *Bull World Health Organ* **2020**, *1*, 32. [CrossRef]
12. Maier, B.F.; Brockmann, D. Effective containment explains sub exponential growth in recent confirmed COVID-19 cases in China. *Science* **2020**, *368*, 742–746. [CrossRef]
13. Dandekar, R.; Barbastathis, G. Neural Network aided quarantine control model estimation of global Covid-19 spread. *arXiv* **2020**, arXiv:2004.02752.
14. Yang, Z.; Zeng, Z.; Wang, K.; Wong, S.-S.; Liang, W.; Zanin, M.; Liu, P.; Cao, X.; Gao, Z.; Mai, Z.; et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *J. Thorac. Dis.* **2020**, *12*, 165–174. [CrossRef] [PubMed]
15. Godio, A.; Pace, F.; Vergnano, A. SEIR Modeling of the Italian Epidemic of SARS-CoV-2 Using Computational Swarm Intelligence. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3535. [CrossRef] [PubMed]
16. Banerjee, A.; Ray, S.; Vorselaars, B.; Kitson, J.; Mamalakis, M.; Weeks, S.; Mackenzie, L.S. Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population. *Int. Immunopharmacol.* **2020**, 106705. [CrossRef]
17. Pinter, G.; Felde, I.; Mosavi, A.; Ghamisi, P.; Gloaguen, R. COVID-19 Pandemic Prediction for Hungary; a Hybrid Machine Learning Approach. *Mathematics* **2020**, *8*, 890. [CrossRef]
18. Efimov, D.; Ushirobira, A. A Prediction of COVID-19 Development in France Based on a Modified SEIR Epidemic Model. Ph.D. Thesis, Inria Lille Nord Europe-Laboratoire CRIStALUniversit´e de Lille, Lille, France, 2020.
19. Breiman, L.; Friedman, J.H. Predicting multivariate responses in multiple linear regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1997**, *59*, 3–54. [CrossRef]
20. Gijbels, I.; Pope, A.; Wand, M.P. Understanding exponential smoothing via kernel regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 39–50. [CrossRef]
21. Exponential Growth Description. 2020. Available online: https://towardsdatascience.com/modeling-exponential-growth-49a2b6f22e1f (accessed on 11 July 2020).
22. Snyder, R.D.; Koehler, A.B.; Hyndman, R.J.; Ord, J. Exponential smoothing models: Means and variances for lead-time demand. *Eur. J. Oper. Res.* **2004**, *158*, 444–455. [CrossRef]
23. Davies, O.L.; Brown, R.G. Statistical forecasting for inventory control. *J. R. Stat. Soc.* **1960**, *123*, 348.
24. Ord, K. Charles Holt's report on exponentially weighted moving averages: An introduction and appreciation. *Int. J. Forecast.* **2004**, *20*, 1–3. [CrossRef]
25. Harvey, A.C. *Forecasting, Structural Time Series Models and the Kalman Filter*; Cambridge University Press: Cambridge, UK, 1990; pp. 25–26.
26. Ng, T.-W.; Turinici, G.; Danchin, A. A double epidemic model for the SARS propagation. *BMC Infect. Dis.* **2003**, *3*, 19. [CrossRef] [PubMed]