Global Journal of Science Frontier Research: F
Mathematics & Decision Science
Volume 20 Issue 5 Version 1.0 Year 2020
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Understanding the Early Evolution of COVID-19 Disease Spread using Mathematical Model and Machine Learning Approaches

By Samuel Oladimeji Sowole, Abdullahi Adinoyi Ibrahim, Daouda Sangare, Ismaila Omeiza Ibrahim & Francis I. Johnson

Baze University

Abstract- In response to the global COVID-19 pandemic, this work aims to understand the early time evolution and the spread of the disease outbreak with a data driven approach. To this effect, we applied Susceptible- Infective-Recovered/Removed (SIR) epidemiological model on the disease. Additionally, we used the Machine Learning linear regression model on the historical COVID-19 data to predict the earlier stage of the disease. The evolution of the disease spread with the Mathematical SIR model and Machine Learning regression model for time series forecasting of the COVID-19 data without, and with lags and trends, was able to capture the early spread of the disease. Consequently, we suggest that if using a more advanced epidemiological model, and sophisticated machine learning regression models on the COVID-19 data, we can understand, as well as predict the long time evolution of the disease spread.

Index Terms: COVID-19, mathematical model, SIR model, machine learning, linear regression, time-series, forecast.

GJSFR-F Classification: *MSC 2010: 93A30* eJournal Print Journal epaper open standard 3d virtual journal

Understanding the Early Evolution of COVID-19 Disease Spread using Mathematical Model and Machine Learning Approaches

Samuel Oladimeji Sowole [α], Abdullahi Adinoyi Ibrahim [σ], Daouda Sangare [ρ],
Ismaila Omeiza Ibrahim [ω] & Francis I. Johnson [¥]

Abstract- In response to the global COVID-19 pandemic, this work aims to understand the early time evolution and the spread of the disease outbreak with a data driven approach. To this effect, we applied Susceptible- Infective-Recovered/Removed (SIR) epidemiological model on the disease. Additionally, we used the Machine Learning linear regression model on the historical COVID-19 data to predict the earlier stage of the disease. The evolution of the disease spread with the Mathematical SIR model and Machine Learning regression model for time series forecasting of the COVID-19 data without, and with lags and trends, was able to capture the early spread of the disease. Consequently, we suggest that if using a more advanced epidemiological model, and sophisticated machine learning regression models on the COVID-19 data, we can understand, as well as predict the long time evolution of the disease spread.

Index Terms: COVID-19, mathematical model, SIR model, machine learning, linear regression, time-series, forecast.

I. Introduction

An outbreak of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) which was first reported in Wuhan, China in December 2019 [1], [2], has led to 3, 910, 738 confirmed cases, 272, 778 death cases, with 2, 352, 811 active cases, and 1, 306, 204 recovered cases as of May 8, 2020, and has spread to 215 countries of the world, [3], [16], [17], [18]. Coronaviruses are a large family of viruses that may cause respiratory illnesses in humans, ranging from common colds to more severe conditions such as Severe Acute Respiratory Syndrome (SARS) and Middle Eastern Respiratory Syndrome (MERS). People that are infected may be sick with the virus for 1 to 14 days before developing any symptoms, [19], [20]. The most common symptoms of coronavirus disease (COVID-19) are fever, tiredness, dry cough, and in severe cases difficulty in breathing [28]. From the statistics given above, it suggests that about 33% of the people infected with the virus will recover from the disease without going through special treatment.

Thus, "Novel Coronavirus" is a new, previously unidentified strain of Coronavirus. This novel coronavirus that is involved in the current outbreak has been named Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) by the International Committee on Taxonomy of Viruses on 11 February, 2020 [28]. While the

Ref

3. *WHO on COVID-19. Available from:, https://www.who.int/emergencies/diseases/novel-coronavirus- 2019,* (Accessed May 2020).

Author α: Department of Mathematical Sciences, African Institute for Mathematical Sciences, Mbour, Senegal.
e-mail: oladimeji.s.sowole@aims-senegal.org
Author σ: Department of Mathematics, Baze University, Abuja, Nigeria. e-mail: abdullahi.ibrahim@bazeuniversity.edu.ng
Author ρ: Department of Mathematics, Universite Gaston Berger, Saint-Louis, Senegal. e-mail: daouda.sangare@ugb.edu.sn
Author ω: Department of Mathematics, University of Maiduguri, Nigeria. email: ibrahimismailaomeiza@gmail.com
Author ¥: School of Architecture and Built Environment, Robert Gordon University, Aberdeen, UK. e-mail: f.johnson6@rgu.ac.uk

disease it causes has been named "Coronavirus Disease 2019" (or "COVID-19" in short) by World Health Organization (WHO) [29]. As a world, we now have a critical situation on our hands with the spread of pandemic COVID-19 which has posed a huge threat to global public health. Economic of the world is also suffering greatly as a result of several measures been put in place including total lock-down of cities in order to curb the spread of the disease. It has also had social, environmental and financial consequences, [18], [21], [22], [23], [30].

Understanding the early transmission dynamics of infectious diseases and evaluating the effectiveness of control measures is crucial for assessing the potential for sustained control when occurring in new areas. Preliminary investigation has suggested that the key influencing factor for the rise in cases across the globe is the disease spread by the infectious people. This means that those with the virus can unknowingly infect others before symptoms appear, some as soon as two days after infection. It is assumed that infected individuals are able to spread the infection until they recover, [24], [25], [26]. As a consequence, several measures have been put in place by the governments and health authorities in order to reduce the spread among the populace. There has been a regular campaign to encourage taking personal measures that will help in reducing the spread. These are the practicing of regular hands-washing, observing social distancing, the use of face masks, disinfection, isolation, among others, [27].

As a new infectious disease, the transmission dynamics of COVID-19 is still under investigation. Al- though SARS-CoV-2 is a kind of coronavirus that is similar to Severe Acute Respiratory Syndrome Coron- avirus (SARS-CoV) and the Middle Eastern Respiratory Syndrome Coronavirus (MERS-CoV), many studies are still underway to understand its infectious characteristics.

Since the outbreak of the disease, several study have been made by researchers to understand the COVID-19 disease, to model and to predict the spread with the help of scientific tools such as mathematical models, predictive analytics using machine learning algorithms, and time series forecasting.

In [9], the author presented a comprehensive under- standing of various pathological mechanism of COVID- 19, which can potentially influence the vulnerable development of the disease. In [10], the author build SARS-CoV-2 computed tomography (CT) scan dataset, consisting of 2482 CT scans for both infected and non-infected patients of SARS-CoV-2 (COVID-19). The data which was collected from Sao Paulo in Brazil is publicly available online to encourage research and implementation of artificial intelligent methods, able to identify an individual infected by COVID-19 through analysis of patient's CT scan. In [11], the authors designed a 3- dimensional structure for 2019-nCoV. To elucidate the 3D structure, a computational modeling approach was applied to give insight into the domain architecture of nsp12. In [12] the authors reviewed the characteristics, origin, pathogenicity, genome structure, and replication of SARS-CoV-2, with the goal of explaining how each remedy strategy could act on slowing down or preventing viral infection. Summary of recently investigated treatments (drugs) were presented. In [13], the authors evaluated the safety and efficacy of 2 chloroquine diphosphate (CQ) dosages in patients with critical SARS-CoV-2. The preliminary findings in this study suggest that higher CQ dosage should not be recommended for patients with critical case of the viral infection because of its potential safety hazards. In [14], a set of novel molecular structures of SARS-CoV-2 3C- like protease inhibitors were presented using the Insilico medicine generative chemistry pipeline. 10 representative structures for likely development with 3D representation were presented. [15] in their work, using the data provided by [6], describes the time-line of a live forecasting exercise, and provides objective forecasts for the confirmed

Ref

23. Cennimo, D. J., & Bronze, M. S. (2020). Coronavirus disease 2019 (COVID-19) treatment & management.

cases of COVID- 19. They discussed also in their work that forecasting of COVID-19 has good potential implications for planning and decision making.

Motivated by [5], [26], [15], the goal of this paper is to study and understand the early-stage evolution of the disease spread with a data-driven approach using Mathematical SIR model and Machine Learning regression model for time series forecasting of the COVID-19 disease, with and without lags and trends.

The rest of the paper is structured as follow: the formulation and application of SIR model to COVID-19 are presented in section II, Data analysis and Time series forecast in section III and section IV give the conclusion to this work.

Ref

15. Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19, PLoS ONE, 15(3): e0231236. https://doi.org/10.1371/journal.pone.0231236.

II. The SIR Model

In this section, we formulate and analyze SIR mathematical model which is the most basic mathematical model for a directly transmitted infectious disease as COVID-19 which is caused by a virus. The transmission of the disease occurs through respiratory droplets, and spread by individual-to-individual contact which happens through a sneeze or cough, through skin-skin contact, or through indirect contact with surfaces in the immediate environment or with objects used on the infected person [4].

a) Mathematical Formulation of SIR Model

The model partitioned the total human population at a particular time into three sub-populations to explains the transmission of the disease in the human population. The total human population which we denoted *N(t)*, is divided into sub-populations of Susceptible class *S(t)*, Infective compartment *I(t)* and Recovered/Removed class *R(t)* So that the total population, *N(t)* is given as:

$$N(t) = S(t) + I(t) + R(t). \tag{1}$$

Susceptible class represent individuals who are at risk of contracting the infection if they had contact with infected individuals. Infective compartment consists of individuals that may not or showing the symptoms of the disease and can infect others. The recovered/removed human compartment are the individuals who have recovered or diseased from the disease. Figure 1 shows flow diagram of COVID-19 disease in a deterministic population. The model variables and their meaning are presented in table I.

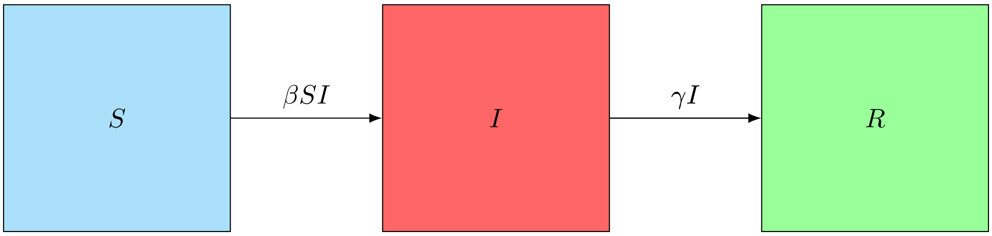

Fig. 1: Flow Diagram of COVID-19 Disease in a Population

Table 1: State variables used and their meanings

Variable	Meaning
$S(t)$	The number of susceptible individuals at a given time, t
$I(t)$	The number of infective individuals at a given time, t
$R(t)$	The number of recovered individuals at a given time, t

The SIR COVID-19 disease transmission model is formulated by assuming the following assumptions as used in [5].

1) There is a large and homogeneously mixing population.
2) The outbreak of the disease is short lived.
3) The model assumed that there is no natural births or natural deaths that occur.
4) The infection has zero latent period. This means that an individual can spread the disease as soon as they become infected with the disease.
5) Recovering from infection may not necessary confers lifetime immunity since there is no vaccine yet to cure the disease.
6) There is an assumption of the mass-action mixing of individuals. By mass action mixing it is assumes that the rate of an encounter between Susceptible and Infective individuals is directly proportional to the product of the population sizes. By this we mean, doubling the size of either Susceptible or Infective population will result in twice as many new infection cases per unit time. This requires that the individuals of both populations are homogeneously distributed in space and thus do mix either in any smaller subgroups or larger. Understandably, every individual will encounter every other individual per unit time with equal probability. Readers should, however, keep it in mind that the SIR model is a deterministic model and there are no probabilities assumptions involved in the formulation of the model.

In the formulation of the model, we do not consider the effect of the natural death or birth rate. This is because the model assumes the effective period of the disease is much shorter than the lifetime of the human. The model lets us know the importance of knowing two parameters, namely β and γ. We will consider also that people develop immunity (in the long term, immunity may be lost and the COVID-19 may come back within a certain seasonality such as also find in the case of common flu disease), and there is no transition from recovered to the remaining two classes.

With the model flow diagram and assumptions above, the differential equations that governed the system are now given below:

$$\frac{dS}{dt} = -\beta SI \tag{2}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{3}$$

$$\frac{dR}{dt} = \gamma I \tag{4}$$

Where β is the contagion rate of the disease and γ is the recovery rate. The disease transmission rate β is > 0 and the recovery rate γ is > 0 also (that is, the duration of infection denoted D is equal $\frac{1}{\gamma}$). The incidence term βSI for the number of new infected individuals per unit time corresponds to homogeneous mixing of the infected and susceptible sub-populations.

It can be easily verified that $\frac{dN}{dt} = 0$, and $N = S+I+R$ is therefore constant. That is, the total population is constant. This easily follows from the SIR system above that the sum of the left hand sides of the three equations is the derivative of the total population size and the sum of the right hand sides is zero. Where the total population size is denoted by N. Since $R(t) = N - S(t) - I(t)$, the system can thus be reduced to a system of two ODEs namely (2) and (3).

Now, Suppose that each infective individual has C contacts that will potentially be from the susceptible class and provided that each infective individual is capable of transmitting the disease per unit time, where C is independent of the population size, then, $C * \frac{S}{N}$ of these contacts are with other susceptible individuals. If the fraction, denoted τ, of adequate contacts result in transmission of the disease, then it follows that each infected individual infects $C * \tau * \frac{S}{N}$ susceptible individuals per unit time. Thus, we have that $\beta = \frac{b}{N}$, where $b = C * \tau$. The parameter τ is called the transmissibility of the COVID-19 disease.

b) Analysis of the SIR Model

1) The Long Term Limits Existence: We show that the long term limits of the SIR model exist.
   Now, since the right hand side of (2) is negative and the right hand side of (4) is positive, this implies that $\frac{ds}{dt} \leq 0$ and $\frac{dR}{dt} \geq 0$.
   Also, since $0 \leq S(t) \leq S(0) \leq N$ and $0 \leq R(0) \leq R(t) \leq N$, this implies that the limits $S(\infty) = \lim_{t\to\infty} S(t)$, $R(\infty) = lim_{t\to\infty} R(t)$, and thus $I(\infty) = \lim_{t\to\infty} I(t) = N - S(\infty) - R(\infty)$ exist.
2) *The Disease Always Dies Out:* It is also easy to prove that the disease always dies out. Now,$I(\infty) = 0$ for all initial conditions, without formulating a formula for $I(t)$. Suppose now that the disease will not dies out, (4) implies that for sufficiently large t, $\frac{dR}{dt} > \frac{\gamma I(\infty)}{2} > 0$ and this implies that $R(\infty) = \infty$, this is a contradiction. Hence, the disease will dies out.
3) *Theorem of Epidemic Threshold:* We define the effective reproductive number, which we denoted $R_e = \frac{S(0)}{N} * \frac{b}{\gamma}$ and the basic reproduction number denoted, $R_0 = \frac{b}{\gamma}$.

Now, if the entire population is initially susceptible, with one infective case, that is, $S(0) = N - 1$, $I(0) = 1$, $R(0) = 0$ and large (see the model assumptions), then $R_e = \frac{N-1}{N} * \frac{b}{\gamma}$ is approximately equal to R_0. Hence, to modify formula involving R_0, we shall assume that the quantity $\frac{N-1}{N}$ is equal to 1. We now show that R_e is the threshold parameter that determines whether the infectious disease will quickly die out or whether it will permeate the population and cause an epidemic.

Theorem 1.

1. *If Re ≤ 1, then I(t) decreases monotonically to zero as $t \to \infty$*
2. *If Re > 1, then I(t) starts increasing, reaches its maximum, and then decreases to zero as $t \to \infty$. The scenario of increasing numbers of infected individuals is called an epidemic. It follows that an infectious disease can get into a population and cause an epidemic in an entirely susceptible population if $R_0 > 1$ or $b > \gamma$.*

Proof.

Equation (3) and the discussion in Section 2.2.1 imply that $\frac{dI}{dt} = (\beta S - \gamma) I \leq (\beta S(0) - \gamma) I = \gamma(R_e - 1) I \leq 0$ for $R_e < 1$. This observation together with I(∞)=0 (see Section 2.2.2) proves the first statement.

Equation (3) implies $(\frac{dI}{dt}(0) = \gamma(R_e - 1)\, I(0) > 0$ for $R_e > 1$. Thus $I(t)$ is increasing at $t = 0$. Equation (3) also implies that I(t) has only one non-zero critical point. These observations, together with $I(\infty) = 0$ imply the second statement is true. This end the proof.

c) Existence and Uniqueness of Solution for the SIR Model

The first-order ODE is generally in the form:

$$x' = f(t, x), \qquad x(t_0) = x_0 \tag{5}$$

In order to established the existence and uniqueness of solution of (5), we are interested in answering the following questions:

(I) Under what conditions can we say solution to equation (5) exists?

(II) Under what conditions can we say there is a unique solution to equation (5)?

In order to answers these questions:
Let:

$$f_1 = -\beta SI,$$

$$f_2 = \beta SI - \gamma I,$$

and

$$f_3 = \gamma I$$

We use the following theorem to established the existence and uniqueness of solution for our SIR model.

Theorem 2 (Uniqueness of Solution)
Let D denotes the domain:

$$|t - t_0| \leq a, ||x - x_0|| \leq b, x = (x_1, x_2, ..., x_n),\ x_0 = (x_{10}, x_{20}, ..., x_{n0}) \tag{6}$$

and suppose that f(t, x) satisfies the Lipschitz condition:

$$||f(t, x_1) - f(t, x_2)|| \leq k||x_1 - x_2||, \tag{7}$$

and whenever the pairs (t, x_1) and (t, x_2) belong to the domain D, where k represents a positive constant.

Then, there exist a constant δ > 0 such that there exists a unique (exactly one) continuous vector solution x(t) of the system (5) in the interval |t − t_0| ≤ δ. It is important to note that condition (2.12) is satisfied by requirement that:

$$\left\{\frac{\partial f_i}{\partial x_j},\ _{i,j=1,2,...,n}\right.$$

be continuous and bounded in the domain D.

Lemma 1: $f(t,x)$ *has continuous partial derivative* $\frac{\partial f_i}{\partial x_j}$ *on a of real bounded closed convex domen* $\mathbb{R}$ *(i.e, convex set of real numbers) where* $\mathbb{R}$ *is used to denotes real numbers, then it satisfies a Lipschitz condition in* $\mathbb{R}$*. Our interest is in the domain:*

$$1 \leq \epsilon \leq \mathbb{R}. \tag{8}$$

So, we look for a bounded solution of the form

$$0 < \mathfrak{R} < \infty.$$

We now prove the following existence theorem.

Theorem 3: (Existence of Solution)

Let D denote the domain defined in (6) such that (7) and (8) hold. Then there exist a solution of model system of equations (2)-(4) which is bounded in the domain D.

Proof. Let:

$$f_1 = -\beta SI, \tag{9}$$

$$f_2 = \beta SI - \gamma I, \tag{10}$$

and

$$f_3 = \gamma I \tag{11}$$

We shows that:
$\frac{\partial f_i}{\partial x_j},\ i, j = 1, 2, 3$ are continuous and bounded. That is, the partial derivatives are continuous and bounded. We explored the following partial derivatives for all the model equations:
From equation (9);

$$\left|\frac{\partial f_1}{\partial S}\right| = \left| - \beta I\right| < \infty,$$

Notes

$$\left|\frac{\partial f_1}{\partial I}\right| = \left| - \beta S\right| < \infty,$$

$$\left|\frac{\partial f_1}{\partial R}\right| = |0| < \infty.$$

Similarly, from equation (10) we also have that:

$$\left|\frac{\partial f_2}{\partial S}\right| = |\beta I| < \infty,$$

$$\left|\frac{\partial f_2}{\partial I}\right| = \left|\beta S - \gamma\right| < \infty,$$

$$\left|\frac{\partial f_2}{\partial R}\right| = |0| < \infty.$$

Finally we have from equation (11);

$$\left|\frac{\partial f_3}{\partial S}\right| = |0| < \infty,$$

$$\left|\frac{\partial f_3}{\partial I}\right| = |\gamma| < \infty,$$

$$\left|\frac{\partial f_3}{\partial R}\right| = |0| < \infty.$$

We have clearly established that all these partial derivatives are continuous and bounded, hence, by Theorem (1), we can say that there exist a unique solution of (9) to (11) in the region D.

d) Implementing the SIR Model Numerically

The SIR model can be implemented in many ways. It can be implemented from the differential equations governing the system, that is within a mean-field approximation. Also, it can be implemented running the system dynamics in a social network (i.e. graph). We will choose the first option and will run a numerical method (with 4th order Runge-Kutta method) to solve the differential equations system.

We define the functions governing the differential equations (2)-(4) using the python programming language. The functions are given below: "

```
# Susceptible equation
def fa(N, a, b, beta):
fa = -beta*a*b
return fa

# Infected equation
def fb(N, a, b, beta, gamma):
fb = beta*a*b - gamma*b
return fb

# Recovered/Deceased equation
def fc(N, b, gamma):
fc = gamma*b
return fc
```

Where $a = S$, $b = I$ and $c = R$.

In order to solve the differential equations system, we develop a 4th order Runge-Kutta (RK4) numerical method using python programming language. Details of RK4 are discussed in [7].

Ref

7. Sowole, S. O., Sangare, D., Ibrahim, A. A., & Paul, I. A. (2019) On the Existence, Uniqueness, Stability of Solution and Numerical Simulations of a Mathematical Model for Measles Disease, *International Journal of Advances in Mathematics*, 2019(4), Pages 84 – 111, 2019.

The results obtained for N = world population, with one initial infected case (that is $I(0) = 1$), $\beta = 0.7$, $\gamma = 0.2$ and a time step $\delta t = 0.1$ are shown in the figure 2.

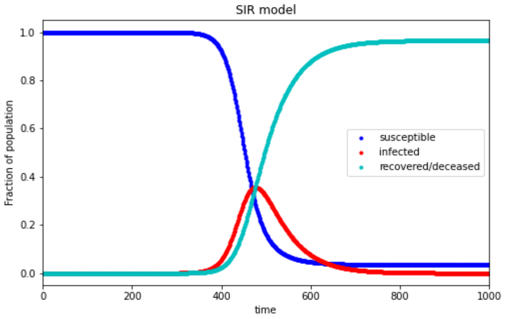

Fig. 2: SIR Model when $\beta = 0.7$ and $\gamma = 0.2$

From the simulation result depicted in the figure 2, the following observations was made:

- The number of infected cases increases for a certain period of time, and then eventually decreases, given that individuals will recovered/deceased from the disease.
- The susceptible compartment of the population de- creases as the virus is transmitted and eventually drop to the absorbent state 0
- However for the recovered/deceased case, the opposite happens.

Readers should however, note that different initial conditions and parameter values will lead to other scenarios from the simulation.

1) *Fitting the SIR model to the Real Data:* Since the beginning of the COVID-19 pandemic, efforts have been put in place to collate live updates of the disease across the globe. For the purpose of our analysis in this paper we are using the data provided by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [6] which provide daily updates on the COVID-19 cases that are curated from different official sources.

The SIR model we modeled relied heavily on theoretical assumptions, and we are interested into real approximation of the COVID-19 expansion in order to extract insights and understand the transmission of the virus. Hence, there is need for us to extrapolate the values of β and γ parameters for each case if we hope to be able to predict the evolution of the system.

We considered selected country of interest and fitted the model with the real COVID-19 data of the country and observed how the evolution of the system looks like with the optimal parameters values of β and γ for the country. See figure 3 below.

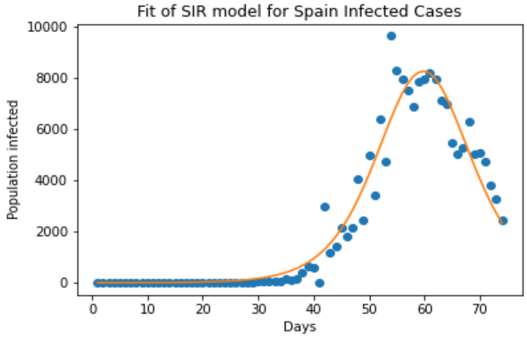

Fig. 3: The Fit of SIR Model for Spain Infected Cases

The optimal parameters extrapolated for the model fit in the figure 3 above are $\gamma = 9.27$ and $\beta = 9.45$. The yellow line indicates the SIR model while the blue dots represent the data points.

III. Data Analysis and Time Series Forecasting of Covid-19 Data

Analyzing the SIR model and the simulations given in the section 2 was meant to understand a model that approximately resembles the transmission mechanism of the COVID-19 virus. However, there are alternative methods that prove to be equally useful both to predict and to understand the pandemic evolution, namely ma- chine learning methods. Many of the popular machine learning methods rely on having rich data to extract conclusions and allow algorithms to extrapolate patterns in data. On the COVID-19 data, we will be considering the machine learning regressions approach under this section.

The COVID-19 dataset we used in the analysis includes dates from January 22 to April 14, 2020, totalling 84 days. This dataset covers 184 countries affected by the disease. Countries with Province/State informed are Australia, Canada, China, Denmark, France, Netherlands, the US, and the United Kingdom. Since the data is close to 3 full months from 2020, this is enough data to get some clues about the pandemic. We give the description of the dataset in Table II below:

Table II: Description of COVID-19 Dataset Used

Feature	Description	Type
Id	Unique ID	Integer
Province_State	Province/State in Country/Region	Character
Country_Region	Country/Region with COVID-19 Cases	Character
Date	Date of case reported	string
ConfirmedCases	Number of Confirmed Cases	float
Fatalities	Number of Fatalities Cases	float

We performed some preprocessing to prepare the dataset for training. This consist of filtering of dates to remove Confirmed Cases and Fatalities post-March 12, 2020 for the short term forecasting purpose, we created additional date columns, analyzed and fixed missing values from the data, and also did data transformations.

Understanding the evolution of the disease spread is vital for the early forecast, so we plot some visuals to understand the evolution of the disease.

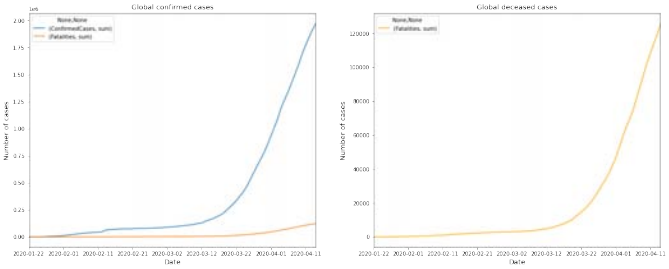

Fig. 4: Evolution of Global COVID-19 Confirmed and Fatalities Cases

Figure 4 show the global COVID-19 curve. It shows a rich fine structure, but these numbers are strongly affected by the vector zero country-China. We called China vector zero country due to the fact that the COVID-19 disease pandemic started there. Now, during the initial expansion of the virus in China, there was no reliable information about the real infected cases. In fact, the criteria that was used to consider as 'infectious cases' in the COVID-19 dataset was modified around February 11, 2020. This strongly perturbed the curve, this is evidence as can seen in the figure 4.

It was observed in figure 5 that the global COVID-19 general behaviour looks cleaner, and in fact, the curve resembles a typical epidemiology model like SIR. We recalled that the SIR mathematical model presents a large increase in the number of infections cases that, once the infectious cases reaches the maximum of the contagion, will then decrease with a lower slope.

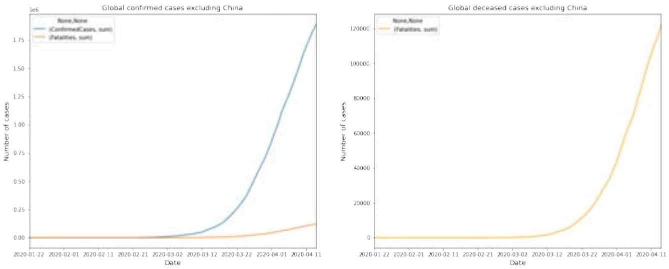

Fig. 5: Evolution of Global COVID-19 Confirmed and Fatalities Cases Excluding China

a) Linear Regression Versus Log-Linear Regression Models

Since we are interested in predicting the future time evolution of the pandemic, our first approach in the paper was to use Machine Learning Linear Regression model for the prediction of the earlier spread of the disease. However, we realized that the evolution is not linear but exponential (this has been discovered from the beginning of the disease spread) so that a preliminary log transformation will be required on the data in order to trained linear model.

The figure 6 below shows a visual comparison of both Confirmed Cases and log Confirmed Cases cases for Spain and with the data information from the last 10 days, beginning from March 1, 2020:

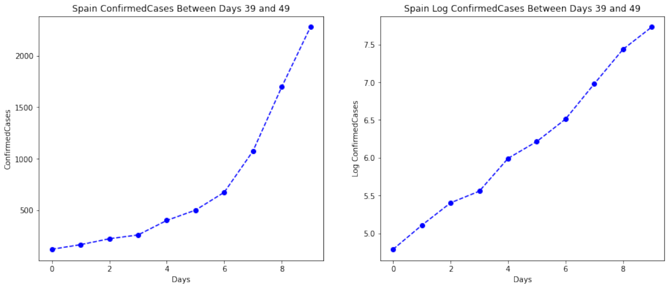

Fig. 6: Spain Confirmed Cases and Log Confirmed Cases

As can be seen in figure 6, the log transformation results in a fancy straight-like line, which is impressive for Linear Regression. However, some important points need to be clarified. This "roughly exponential behaviour" of the curve is only true for the initial disease spread of the COVID-19 pandemic (that is the initial increase of infections on the SIR model). But that is exactly the point where most countries are at as at the time of writing this paper. In our analysis, we only extract the last 10 days of data in order to capture exactly the very short term component of the evolution to prevent the effects of certain variables that have been impacting the transmission speed, for instance, the effect of quarantine in comparison to when there is free circulation, etc, and also to prevent differences on criteria when confirming new cases.

We follow the steps below in training the linear regression model:

- Features Selection. We select features for the training.
- Dates filtering. We filtered the train data from March 1 to March 18, 2020.
- Application of Log transformation. We applied log transformation to Confirmed Cases and Fatalities features.
- Handled infinities. We Replace infinities from the logarithm with 0 values. Given the asymptotic behaviour of the logarithm for log(0), this implies that when applying the inverse transformation (exponential) a 1 will be returned instead of a 0. This problem does not impact many countries. In our subsequent work, we intend to address this issue in order to obtain a cleaner solution.
- train/test split. Split the data into train/validation/test sets
- Prediction. Apply Linear Regression, and trained the model country by country after which the results was joined together.

Figures 7 to 10 shows the results for some of the countries after training the model.

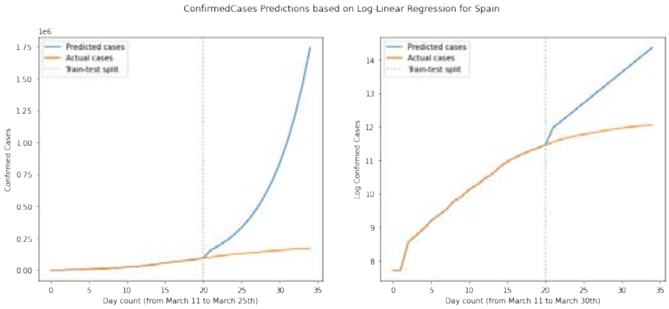

Fig. 7: 1st Predictions for Confirmed Cases based on Log- Linear Regression for Spain

From the regression results, some of which were depicted in the figure 7 to 10, the following general observations were noticed:

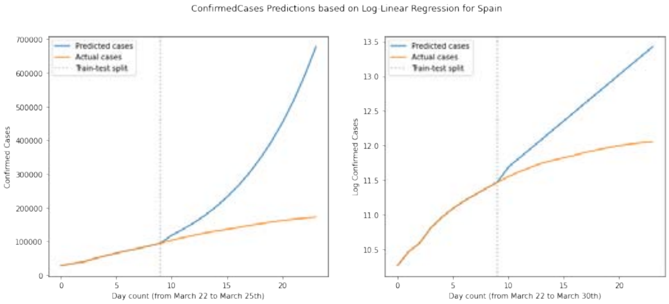

Fig. 8: 2nd Predictions for Confirmed Cases based on Log-Linear Regression for Spain.

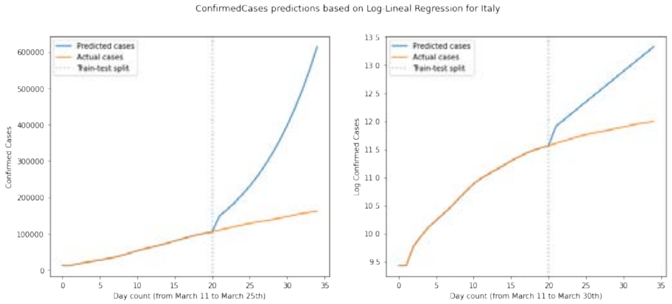

Fig. 9: 1st Predictions for Confirmed Cases based on Log- Linear Regression for Italy

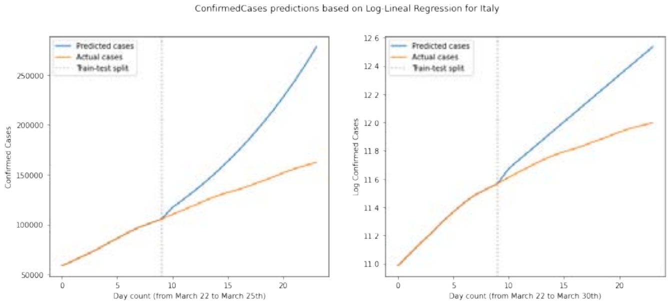

Fig. 10: 2nd Predictions for Confirmed Cases based on Log-Linear Regression for Italy

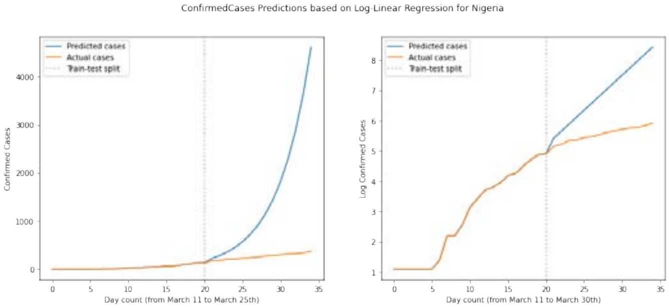

Fig. 11: 1st Predictions for Confirmed Cases based on Log-Linear Regression for Nigeria

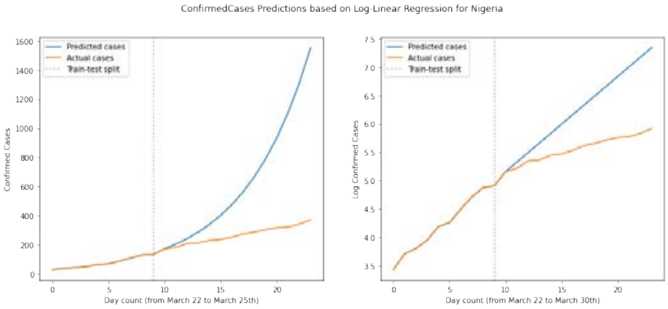

Fig. 12: 2nd Predictions for Confirmed Cases based on Log-Linear Regression for Nigeria

- The general evolution of the disease spread is captured despite the simplicity of the model.

- The cumulative infected cases have been changing since March, so that using the whole month data for training the model will result in overestimated predictions. When we reduced the training set to only a few days prior to the testing region, results are better. This is capturing the problem of the exponential behaviour that is only true for the early stages of the disease spread. The stage we are now, the disease spread behaviour is more complex, and in order to predict the evolution with large portions of the historic data, alternative and better models will be required (e.g sigmoid, ARIMA, etc.).
- Estimations are increasingly worse as time passes (getting harder to extrapolate).
- Countries that recently confirmed their first contagions are difficult to predict (i.e, countries with fewer data points)
- Countries with 0 cases in the whole training dataset are predicted as non-infected (countries with no data points).

b) Linear regression with Lags and Trends

With the results obtained in the previous subsection, we can deduce that the Linear Regression Model is a good model for predicting the early stages of the COVID-19 spread. That is, Linear regression model can predict the initial outbreak of the disease from the data we are analyzing, and there's no way our model could predict when the number of new infections is going to decrease. But for short-term prediction purposes, everything is fine. We will now try to improve the results obtained previously in this subsection.

Time series data possess specific properties such as trend and structural break, common methods used to analyze other types of data may not be appropriate for the analysis of time series data [8]. So, enriching time series data is key to obtain good results, therefore, we applied two different transformations to the COVID-19 data in this subsection, namely lags and trends. For example, on Confirmed Cases column, lags can be seen as a way to compute the previous value of the column so that the lag_3 for Confirmed Cases would inform this column from the previous day. By definition, the lag_3 of a feature X is given by the formula:

$$X_{lag_3}(t) = X(t-3) \qquad (12)$$

Transforming a column into its trend gives the natural tendency of this column, which is different from the raw value. The definition of trend we applied is given by:

$$Trend_X = \frac{X(t) - X(t-1)}{X(t-1)} \qquad (13)$$

The backlog of Lags we will apply is 14 days while for Trends is 7 days both for Confirmed Cases and Fatalities.

In order to apply lags in the model, there is a problem to solve: If we use our dataset to predict the next following days of contagions, for the first day, all the lags will be reported (from the previous days), but for the next days, many of the lags will be unknown (and will be flagged as 0), since the number of Confirmed Cases is only known for the training subset. We take the following simple approach to overcome this problem:

(i) We begin with the train dataset, with all cases and lags reported

Ref

8. Min B. Shresthaa, Guna R. Bhatta. (2018) Selecting appropriate methodological framework for time series data analysis, *The Journal of Finance and Data Science*, 2019(4), Pages 71 – 89, 2018.

(ii) We forecast only for the next day through the Linear Regression

(iii) We set the new prediction as a Confirmed Cases and Fatalities

(iv) We recompute lags

(v) We then repeat from step (ii) to step (iv) for all the remaining days.

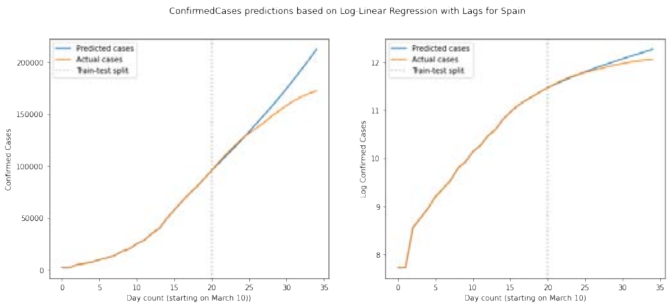

Fig. 13: Predictions for Confirmed Cases of Log-Linear Regression with Lags for Spain

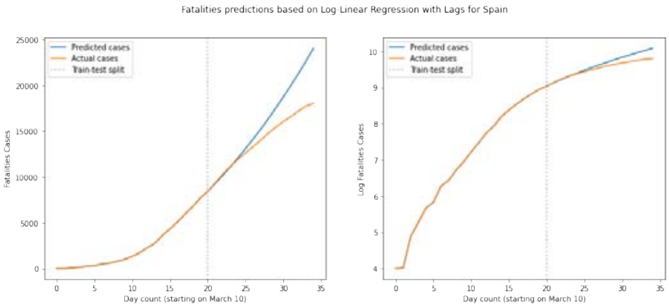

Fig. 14: Predictions of Fatalities of Log-Linear Regression with Lags for Spain

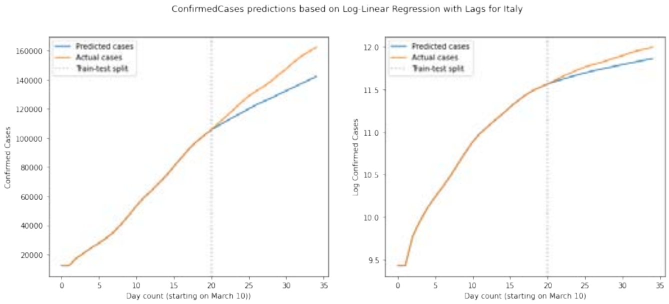

Fig. 15: Predictions for Confirmed Cases of Linear Regression with Lags for Italy

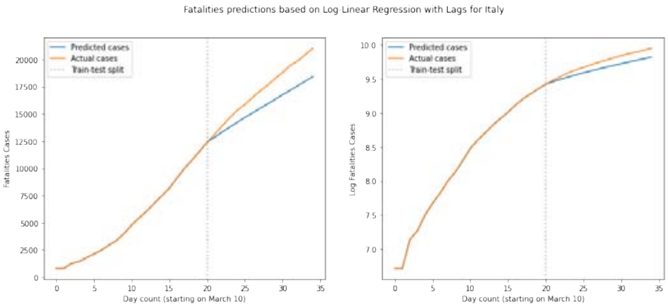

Fig. 16: Predictions of Fatalities of Linear Regression with Lags for Italy

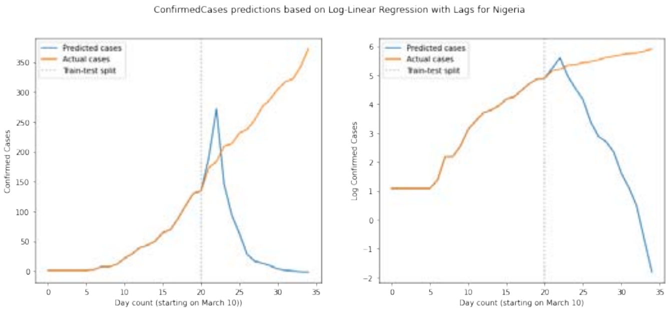

Fig. 17: Predictions for Confirmed Cases of Linear Regression with Lags for Nigeria

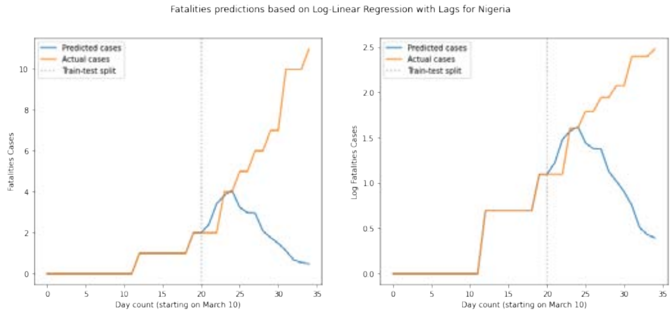

Fig. 18: Predictions of Fatalities of Linear Regression with Lags for Nigeria

Figure 13 to 18 shows some of the regression results with the addition of lags. Two full weeks of the data was used for the training (from February 26th to March 11th), with their previous 14 lags. With the addition of the lags, the following observations were made:

1. *Countries with enough data:* For countries with several Confirmed Cases and Fatalities (For ex- ample, Spain, Italy) in the train dataset (prior to March 11th), predictions are very precise and similar to actual data points.
2. *Countries with poor data:* Countries with a small number of data points (e.g. Nigeria) in the train dataset show a potentially disastrous prediction. Given the small number of cases, the log transformation followed by a Linear Regression is not able to capture the future behavior of the disease spread.
3. *Countries with no data:* When the number of con- firmed cases in the train dataset is 0 or negligible, the model predicts always no infections.

IV. Conclusion

Understanding the evolution of the spread of a disease like COVID-19 will ultimately play a major role in prevention measure that can be taken by relevant authorities to flatten the curve. The aim of this paper was to explore and understand the early time evolution of the spread of COVID-19 disease with a data-driven approach. Therefore, we modelled and applied COVID- 19 disease to the SIR epidemiological model. Also, we use the machine learning linear regression model on the COVID-19 data that is publicly available for early-stage time series forecasting of the COVID-19 cases, considering when there is no lags and trends, and with the application of lags and trends. With these approaches, we are able to capture and understand the early spread of the disease. Hence, we suggest that if using a more advanced deterministic epidemiological model, and sophisticated machine learning regression models on the COVID-19 data, it is possible to understand, as well as predict the long time evolution of the disease spread.

Declaration of Interests

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

Funding Source

The authors declare that there is no funding received for the writing of this paper.

Notes

References Références Referencias

1. *Everything You should Know about the Coronavirus outbreak. Available from:, https://www.pharmaceutical-journal.com/news-and-analysis/features/everything-you-should-know-about-the-coronavirus-outbreak/20207629.article?firstPass=false,* (Accessed May 2020).
2. Muhammad Adnan Shereena, Suliman Khana, Abeer Kazmic, Nadia Bashira, Rabeea Siddiquea,. (2020) COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses, *Journal of Advance Research,* 2020(24), Pages 91−98, 2020.
3. *WHO on COVID-19. Available from:, https://www.who.int/emergencies/diseases /novel-coronavirus- 2019,* (Accessed May 2020).
4. *Transmission Mode of COVID-19. Available from:, https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implica-tions-for-ipc-precaution-recommendations,* (Accessed May 2020).
5. *The SIR model and the Foundations of Public Health. Available from:, http://mat.uab.cat/matmat/PDFv2013/v2013n03.pdf,* (Accessed May 2020).
6. *Johns Hopkins COVID-19 Datasets. Available from:, https://github.com/CSSEGIS and Data/COVID-19,* (Accessed May 2020).
7. Sowole, S. O., Sangare, D., Ibrahim, A. A., & Paul, I. A. (2019) On the Existence, Uniqueness, Stability of Solution and Numerical Simulations of a Mathematical Model for Measles Disease, *International Journal of Advances in Mathematics,* 2019(4), Pages 84 − 111, 2019.
8. Min B. Shresthaa, Guna R. Bhatta. (2018) Selecting appropriate methodological framework for time series data analysis, *The Journal of Finance and Data Science,* 2019(4), Pages 71 − 89, 2018.
9. Kowalik, M.K. (2020) COVID-19 - Toward a Comprehensive Understanding of the Disease., *Cardiology Journal,* 2020.
10. Soares, E. A. (2020) SARS-CoV-2 CT-Scan Dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification, *medRxiv,* 2020.
11. Abbas Khan, M. K. (2020) Decoding the structure of RNA- dependent RNA polymerase (RdRp), understanding the ancestral relationship and dispersion pattern of 2019 Wuhan Coronavirus, *PREPRINT (Version 1) available at Researchgate,* April 29, 2020.
12. Alanagreh, L. A. (2020) The Human Coronavirus Disease COVID-19: Its Origin, Characteristics, and Insights into Potential Drugs and Its Mechanisms, *Pathogens,* 9(5), 331, 2020.

13. Borba, M. G., (2020), Effect of high vs low doses of chloro- quine diphosphate as adjunctive therapy for patients hospitalized with severe acute respiratory syndrome coronavirus 2 (SARS- CoV-2) infection: a randomized clinical trial, *JAMA Network Open,* 3(4), e208857, 2020.
14. Zhavoronkov, A. A., (2020) Potential COVID-2019 3c-like protease inhibitors designed using generative deep learning ap- proaches, *Insilico Medicine Hong Kong Ltd A*, 307, E1.
15. Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19, PLoS ONE, 15(3): e0231236. https://doi.org/10.1371/journal.pone.0231236.
16. *COVID-19 Live Dashboard. Available from:, https://sowole- etal-covid-19-dashboard.herokuapp.com/,* (Accessed May 2020).
17. Pal, M., Berhanu, G., Desalegn, C., & Kandi, V. (2020). Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): An update. Cureus, 12(3).
18. Park, S. E. (2020). Epidemiology, virology, and clinical features of severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2; Coronavirus Disease-19). Clinical and Experimental Pedi- atrics, 63(4), 119.
19. Paules, C. I., Marston, H. D., & Fauci, A. S. (2020). Coro- navirus infections—more than just the common cold. Jama, 323(8), 707-708.
20. Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J., & Yan, Y. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. Military Medical Research, 7(1), 1-10.
21. Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2020). Features, evaluation and treatment coronavirus (COVID-19). In Statpearls [internet]. StatPearls Publishing.
22. Rabi, F. A., Al Zoubi, M. S., Kasasbeh, G. A., Salameh, D. M., & Al-Nasser, A. D. (2020). SARS-CoV-2 and coronavirus disease 2019: what we know so far. Pathogens, 9(3), 231.
23. Cennimo, D. J., & Bronze, M. S. (2020). Coronavirus disease 2019 (COVID-19) treatment & management.
24. World Health Organization, & World Health Organization. (2020). Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19).
25. ASSESSMENT, R. R. (2020). Coronavirus disease 2019 (COVID-19) in the EU/EEA and the UK–ninth update.
26. Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Ed- munds, J., Funk, S., ... & Davies, N. (2020). Early dynamics of transmission and control of COVID-19: a mathematical mod- elling study. The lancet infectious diseases, 20(5), Pages 553 - 558, 2020.
27. World Health Organization. (2020). Overview of public health and social measures in the context of COVID-19: in- terim guidance, 18 May 2020 (No. WHO/2019-nCoV/PHSM- Overview/2020.1). World Health Organization.
28. Muhammad Adnan Shereena, Suliman Khana, Abeer Kazmic, Nadia Bashira, Rabeea Siddiquea,. (2020) COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses, Journal of Advance Research, 2020(24), Pages 91 - 98, 2020.
29. J. Cui, F. Li, Z.-L. Shi,. (2019) Origin and evolution of pathogenic coronaviruses, Nat Rev Microbiol, 2019(17(3)), Pages 181 - 192, 2019.

30. Peterson K. Ozili, Thankom Gopinath Arun. (2020) Spillover of COVID-19: impact on the Global Economy, SSRN Electronic Journal, DOI: 10.2139/ssrn.3562570, March, 2020.

Notes