Chapter

How Can We Be Ahead of COVID-19 Curve? A Hybrid Knowledge-Based and Modified Regression Analysis Approach for COVID-19 Tracking in USA

Rafaat Hussein

Abstract

Since its appearance in 2019, the COVID-19 virus deluged the world with unprecedented data in short time. Despite the countless worldwide pertinent studies and advanced technologies, the spread has been neither contained nor defeated. In fact, there is a recent record surge in the number of confirmed new cases. The rational question is thus: why has it taken so long to date to forecast the trajectory of the spread? To this end, this chapter presents a new predictive Knowledge-based (KB) toolkit named CORVITT (Corona Virus Tracking Toolkit) and a modified linear regression model. This logical step assists the officials, organizations, and users to forecast the spread trajectory and accordingly make proactive rather than retroactive intervention decisions. This hybrid approach uses the confirmed new cases and demographic data, implemented. CORVITT is not an epidemiological model, in the sense that it does not model disease transmission, nor does it use underlying epidemiological parameters or data including the reproductive rate, disease methods, real time polymerase chain reaction cycle threshold, the virus structure and pathogenesis, etc. The chapter is a seed in an in-progress study that will broaden its scope by including additional parameters.

Keywords: COVID-19, coronavirus, epidemic, modelling, outbreak, pandemic

1. Introduction

The unfolding COVID-19 has turned the world upside down [1], and this unprecedented trend is set to be the worst pandemic of a generation in terms of the increasing number of infected people. In its report on April 7, 2020, the US Centers for Disease Control and Prevention (CDC) [2] indicated that the COVID-19 poses a severe threat to public health. In its report, the CDC indicated that the "complete clinical picture with regard to COVID-19 is not fully known." To deal with this blurred picture, the World Health Organization (WHO) has compiled an overwhelmingly pertinent database [3]. The CDC provides a daily report that includes new data reported to the CDC by 55 USA jurisdictions [4]. Many other organizations have also provided similar resources and statistics including the Chinese

Medical Association Publishing House [5] and the European Centre for Disease Prevention and Control [6]. All the available approaches suggest that the number of new COVID-19 cases plays a key role in mapping its trajectory [7] worldwide.

COVID-19 is an evolving epidemic, and its up-and-down spread (trend or pattern commonly referred to as "curve") is a sign of its elusiveness. As of today (July 25, 2020), the COVID-19 is striking back with record-setting blows. In general, the COVID-19 issue relates to various facets such as public health and social as well as culture characteristics, and the world seems lacking sound methodologies on how to address this problem. Using predictive tracking or forecasting quantitative measures can assist the authorities, officials, organizations, and users to be proactive rather than reactive, and thus better prepared to mitigate potential adversaries.

2. The model

The literature seems to suggest that using the number of new cases and the level of social distancing are the key variables to analyze the COVID-19 in various ways. In what follows, we provide a background information about the four main COVID-19 modeling techniques: system dynamics, agent-based modeling, discrete event simulation, and hybrid simulation [8]. System dynamics uses differential equations to model resources, knowledge, people, and money, and the flows between these parameters explains the simulation behavior. The agent-based techniques are stochastic, enabling the variability of human behavior to be incorporated to help understand the likely effectiveness of proposed protective measures. The discrete event technique is also stochastic and models operations over time where entities flow through a number of activities. The hybrid simulation combines two or more techniques and is used for complex behavior. These techniques focus mainly on the unfolding phases of disease transmission such as quarantine, lock down, testing, and health care services. Some of these approaches have been rooted in the literature since 1777, and are complex, and cumbersome to implement. Without adequate specialists in advanced and complex mathematical theories and/or computers, the logical question is thus: how could the proper personnel ascertain the COVID-19 spread in order to make proactive intervention decisions; e.g. to prepare hospitals and intensive care units, to mitigate the adverse impacts of what may happen in the near future? In search for accurate answer and based on the popular utilization of COVID-19 relationship between the number of cases [9] and population per land area, the idea of a new index was conceptualized in this study. It represents the number of reported confirmed new cases per population in the specific region the data was recorded. This new concept harnesses the number of cases and the regional crowdedness of people, which varies in the US from single digit to multi-thousand [2]. The index increases with more cases and with more dense populations (assumed shorter social distancing).

In this study, a combined linear regression analysis and data-fitting model is used. To deal with data fluctuation, this model adopted the hypothesis that was successfully used in other published studies of a short time span of one month maximum for forecasting, [10–13]. That hypothesis is logical and rational because the world knows that the virus spread in unpredictable; thus, longer time spans may encompass inaccurate data. The data is obtained from the New York Times Journal database [14]. The journal publishes the daily cases of COVID-19 by state and county in the US. The data from eleven states was used: New York State (NYS), Florida (FL), California (CA), Colorado (CO), Illinois (IL), Texas (Tx), Louisiana (LA), Washington (WA), Georgia (GA), New Jersey (NJ), and Michigan (MI).

We first constructed the slope for the confirmed cases and population in each of the states from March 27 to May 11, 2020, and then used polynomials fitting and linear regression analysis for forecasting. Linear regression is a direct way to deal with the connections between variables.

3. New knowledge-based toolkit

To accurately and proactively capture the big picture of COVID-19 spread, this study transfers the expertise of problem-solving from humans into a KB toolkit that takes in the same data, and yields the same conclusion but faster. This new KB-statistic hybrid approach effectively assists humans in dealing with COVID-19 massive daily data in addition to save time which is an essential requirement in dealing with the virus illusiveness. The study introduces for the first time in this field, to our knowledge, a novel KB toolkit to visualize the data and make it easier to understand and use without either mathematical or computer expertise. The CORVITT is a promising incubator for COVID-19 future forecasting platforms. Its VBA-based architecture blueprint emerges from an open-end modular adaptable structure encompassing a graphical-interface client allowing the users to easily operate it. This KB technology has been proven in other applications and thus applied in this study for COVID-19 [15–17]. To the author's knowledge, the concept of CORVITT has not been attempted to date for COVID-19. **Figure 1** shows the dashboard of CORVITT. The user could simply click the button that represents the state/province of interest, and the dashboard will display the microdata or the relative comparison of all states. **Figure 2** shows the data used in **Figure 1**. Although the amount of collected data is massive, the use of the dashboard is intuitive and user

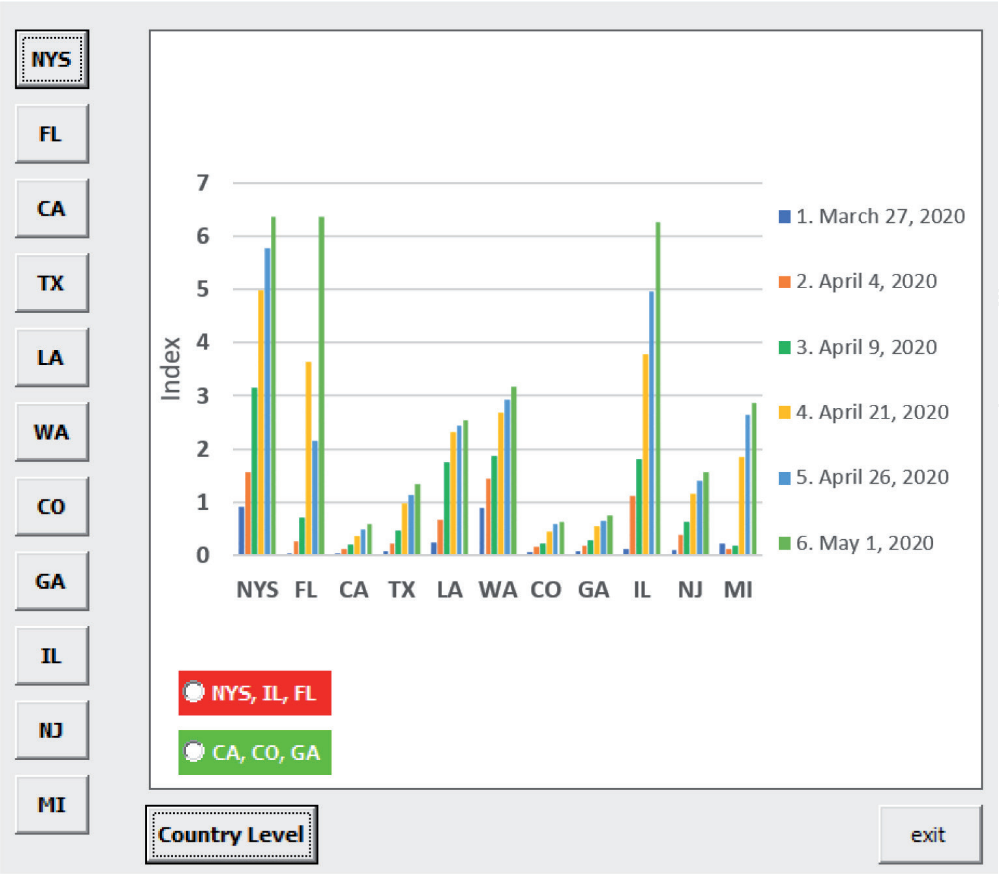

Figure 1.
Dashboard of the CORVITT presented in this chapter.

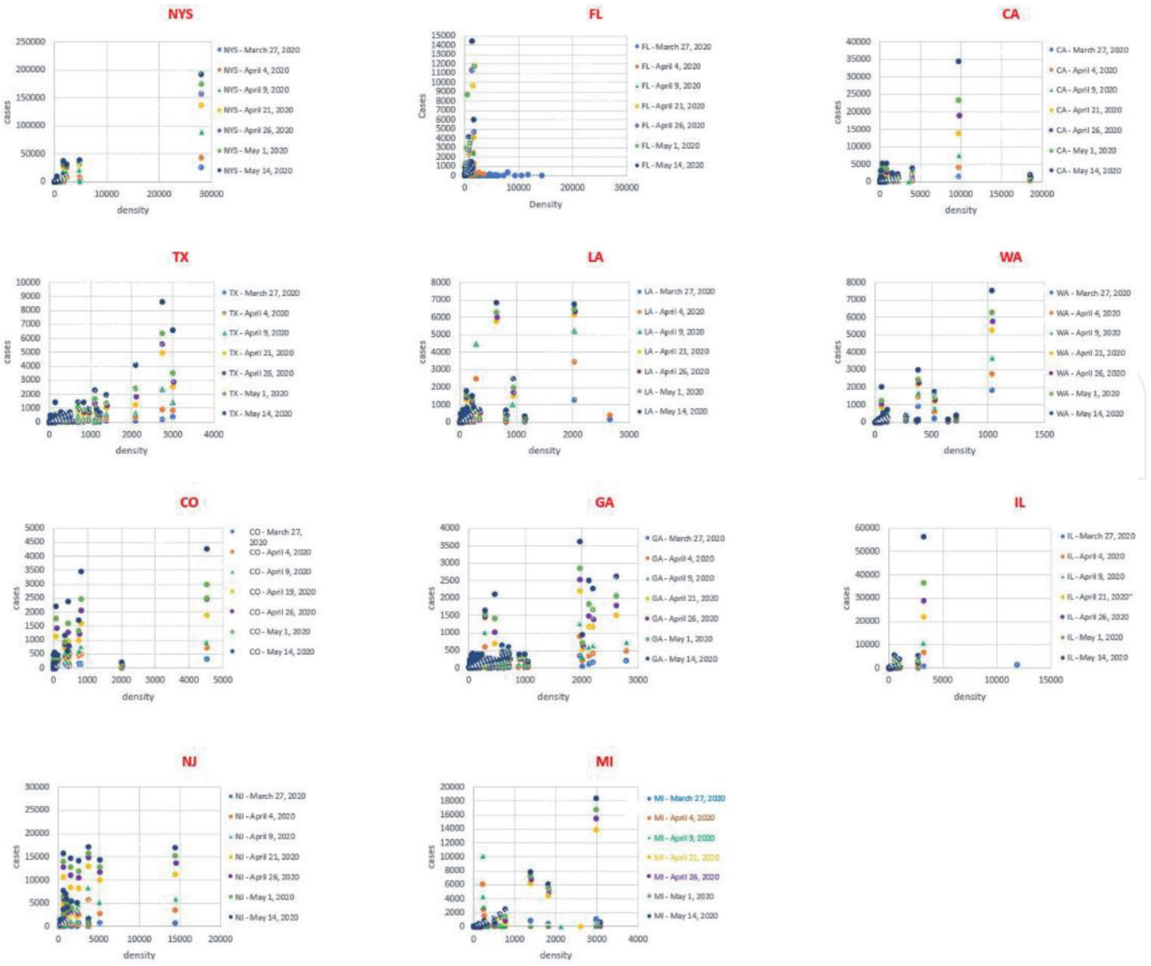

Figure 2.
The macro-data used in this study.

friendly. Again, there is no need for medical, mathematical, or computer skills to use the dashboard and benefit from its applications. This is one of the takeaways of bringing an artificial brain to help human brains in dealing with complex challenge at hand such as the COVID-19.

4. Results and discussion

In what follows, we examined the feasibility of the ascribed model for COVID-19 in two ways: firstly, by analyzing its forecasted outcomes in eleven US states, and secondly by comparing the forecasted results with actual onsite data. Firstly, a database was created at micro-level or counties, for the first time to our knowledge, for the new cases and population per land area on COVID-19 from March 27 to May 1, 2020 in the eleven US states: NYS, FL, CA, TX, LA, WA, CO, GA, IL, NJ, and MI. These sates had a steady high number of confirmed cases according to the New York Times Journal. **Table 1** shows some of the collected data. **Figure 2** shows CORVITT gauging of the virus county-wise distributions in terms of the new index and population. **Figure 2** shows that in NYS, FL, CA, CO, and IL the inhabitants are infected in areas with large social distances because most of the data is concentrated at low population, i.e. large spaces between the inhabitants. On the contrary, in TX, LA, WA, GA, NJ, and MI the virus was spread though the social distance was small because the data spreads over a wide range of population, i.e. large space between the inhabitants. Unlike the common general approach that was used for all US states

Regression statistics						
Multiple R	0.992092					
R Square	0.984246					
Adjusted R Square	0.976369					
Standard error	0.372065					
Observations	4					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	17.29721	17.29721	124.9503	0.007908	
Residual	2	0.276865	0.138433			
Total	3	17.57407				
	Coefficients	**Standard error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95.0%**
Intercept	−7476.32	669.1315	−11.1732	0.007915	−10355.4	−4597.28
date	0.170252	0.015231	11.17812	0.007908	0.104719	0.235785

Table 1.
A sample output of the modified regression.

at all times since 2019, which is common in the news media from Tabloid to New York Times journals, this discovery unveiled new facets. **Figure 1(d)** shows that on March 27, the indexes were 0.93 and 0.10 in NYS and NJ (large distance), respectively, 0.08 and 0.07 for TX and GA (small social distance), respectively, though the spread of data appeared similar on the dashboard in each group. In addition, NYS, LA, WA, IL, and MI have high indexes by comparison to FL, CA, TX, CO, GA, and NJ For example, on April 21, 2020, the indexes were 5.0 and 0.40 for NYS and CA although the closeness of data in both states was similar. Furthermore, the increase of the index within each state was nonuniform. For example, the index increased from 2.8 to 5 between April 9 and 21 in NYS, and from 0.05 to 1.0 in TX over the same period. From a different angel of view, **Figure 1(a)** shows that the rate of spreading of the virus differ from one state to another. For example, the spreading rate in IL is very high compared to that in CA. This indicates that the virus can affect more people in IL (57,920 sq. mi and 13 million people) than NJ (8730 sq. mi and 9 million people). Taken as a whole, CORVITT-outcomes suggest that NYS is a good region for the virus to spread whereas CA is not as good from March 27 to May 1, 2020. Such information allows the authorities to prioritize the resources giving NYS the highest priority. Secondly, **Figure 3a** and **b** compares the forecasted and actual on-site data and shows a close agreement in different US states. **Figure 3a** shows that whereas the new cases in NYS has reached the peak in the first week of May, the pandemic was worsening in other states such as Illinois, but in California, Georgia, and Colorado reached a plateau. **Figure 3c** shows the severity of the pandemic as indicated by the skewness; positive (to the right) for LA and negative for NYS, NJ, and MI of the growth and deterioration of the distributions [18]. The positive skewness means longer deterioration (decline in cases) time. The shallow deterioration rate at the trailing end of the curve in **Figure 1(b)** is a sign of a plateau. **Figure 1** describes the peak, weakness, and steadiness statuses by which the virus trajectory disperses through different stages in various regions. This new discovery is useful to understand the building up and collapse of the virus impacts thus make proactive preparations possible.

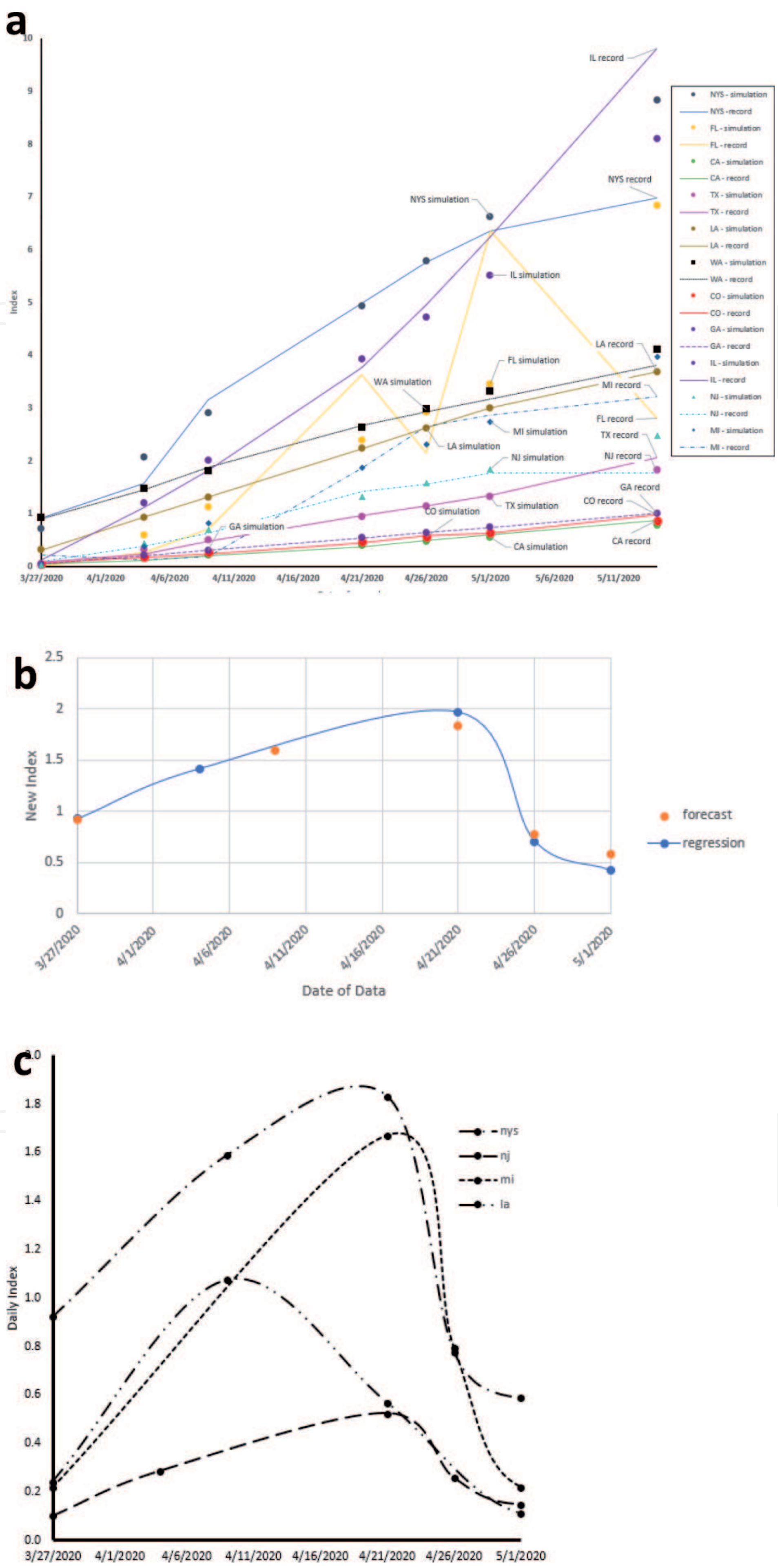

Figure 3.
(a): The trajectory of cases' dispersion over time in various US States. (b): A satisfactory agreement between the forecasted and actual data. (c). The growth and deterioration distributions of cases over time.

5. Conclusions

As the enormity of the COVID-19 threat has become clear, the characteristics of existing COVID-19 complex analytic methodologies and the all-encompassing approach place serious limitations on their usefulness for practical use. The computer technologies have reached what no one could imagined, and the KB systems have proven very beneficial in many fields. The rational question is: why has it taken so long for a logical approach to appear to practicalize the analytical complex simulations? To answer the question, this chapter introduces machine smartness to assist humans' intelligence to capture the big picture of the virus illusiveness thus take proactive rather than retroactive steps to mitigate safely its inevitable adverse effects. This seed study introduced a hybrid KB-regression analysis model for COVID-19 forecasting. It used data collected from eleven US states at macro-level level to foresee the short-term spread trajectory. The outputs unveiled new discoveries and shed light on various facets of the COVID-19 in each state. The accuracy of the hybrid approach was gauged by comparing forecasted and actual data and satisfactory agreements were found. It should be noted that this study is a step forward, but additional development is in progress for improvement preparations.

Author details

Rafaat Hussein
State University of New York, Syracuse NY, US

*Address all correspondence to: ezpsc@yahoo.com

References

[1] United nations Committee for the Coordination of Statistical Activities, How covid-19 is changing the world, 2020. https://unstats.un.org/unsd/ccsa/documents/covid19-report-ccsa.pdf.

[2] The Centers for Disease Control and Prevention [CDC], Situation Summary, 2020. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/summary.html.

[3] The World Health Organization [WHO], Global research on coronavirus disease [covid-19], 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov.

[4] The Centers for Disease Control and Prevention [CDC], Cases in U.S., 2020. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html.

[5] The Chinese Medical Association Publishing House, COVID-19 Academic Research Platform, 2020. http://eng.med.wanfangdata.com.cn/Chinese_Medical_Association_Publishing_House.html.

[6] The European Centre for Disease Prevention and Control, Coronavirus disease, 2020. https://www.ecdc.europa.eu/en.

[7] STAT, The coronavirus is washing over the U.S. These factors will determine how bad it gets in each community, 2020. https://www.statnews.com/2020/04/01/coronavirus-how-bad-it-gets-different-communities/

[8] Currie, C. et al., How simulation modelling can help reduce the impact of covid-19, Journal of Simulation, Vol. 14, pp: 83-97, 2020. https://www.tandfonline.com/doi/pdf/10.1080/17477778.2020.1751570?needAccess=true

[9] Our World in Data, COVID-19 death rate vs. Population density, 2020. https://ourworldindata.org/grapher/covid-19-death-rate-vs-population-density

[10] Yousaf, M. et al., Statistical analysis of forecasting covid-19 for upcoming month in Pakistan, Elsevier Public Health Emergency Collection, 2020. DOI: 10.1016/j.chaos.2020.109926

[11] Sujath, R. et al., A machine learning forecasting model for covid-19 pandemic in India, Stochastic Environmental Research and Risk Assessment, 2020. https://doi.org/10.1007/s00477-020-01843-8.

[12] Yonar H. et al., Modeling and Forecasting for the number of cases of the covid-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods, Eurasian Journal of medicine and Oncology, pp: 160-165, 2020. DOI: 10.14744/ejmo.2020.28273

[13] Öztoprak,F. and Javed, A, Case Fatality Rate estimation of covid-19 for European Countries: Turkey's Current Scenario Amidst a Global Pandemic; Comparison of Outbreaks with European Countries, Eurasian Journal of medicine and Oncology, Vol. 4 No. 2, pp: 149-159, 2020. DOI: 10.14744/ejmo.2020.60998

[14] The New York Times Journal, The Coronavirus in the US, 2020. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

[15] Hussein, R., Knowledge-based & expert system technologies for built & natural environments, 2020, https://aitechm.wixsite.com/8kbextech.

[16] Hussein, R., Artificial Intelligence for the Built and Natural Environments

and Climate, In: Advances in Artificial Intelligence, edited by Sergey Yurish, the International Frequency Sensor Association, 2019, Chapter 7.

[17] Hussein, R., Knowledge-based Tools for Monitoring and Management of the Engineered Infrastructure Construction Systems, In: Advances in Computers and Software Engineering, edited by Sergey Yurish, International Frequency Sensor Association, 2020, Chapter 5.

[18] U.S. Department of Health and Human Services, Principles of Epidemiology in Public Heath Practice, 2012. https://www.cdc.gov/csels/dsepd/ss1978/index.html