Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Prediction model for the spread of the COVID-19 outbreak in the global environment

Ron S. Hirschprung [*], Chen Hajaj

Department of Industrial Engineering and Management, Ariel University, Israel

ARTICLE INFO

Keywords:
COVID-19
Data mining
Decision support systems
Machine learning
Prediction methods

ABSTRACT

COVID-19 has long become a worldwide pandemic. It is responsible for the death of over two million people and posed an economic recession. This paper studies the spread pattern of COVID-19, aiming to establish a prediction model for this event. We harness Data Mining and Machine Learning methodologies to train regression models to predict the number of confirmed cases in a spatial-temporal space. We introduce an innovative concept – the Center of Infection Mass (CoIM) – adapted from the field of physics. We empirically evaluated our model on western European countries, based on the CoIM index and other features, and showed that a relatively high accurate prediction of the spread can be obtained. Our contribution is twofold: first, we introduced a prediction methodology and proved empirically that a prediction can be made even to the range of over a month; second, we showed promise in adopting the CoIM index to prediction models, when models that adopt the CoIM yield significantly better results than those that discard it. By applying our model, and better controlling the inherent tradeoff between life-saving and economy, we believe that decision-makers can take close to optimal measures. Thus, this methodology may contribute to public welfare.

1. Introduction

The novel Coronavirus (COVID-19), which was declared a Public Health Emergency of International Concern (PHEIC) by the World Health Organization (WHO), started locally in a wholesale fish and seafood market in Wuhan, China, spread very quickly (Zu et al., 2020), and has become a worldwide pandemic (WHO & others, 2020). The WHO warned that “2 Million Coronavirus Deaths Is Not Impossible” (CNBC, 2020), sadly, a forecast that came true. COVID-19 has now affected hundreds of countries, as of late January 2021 is responsible for the death of over two million people, and on average, the death rate per number of diagnosed cases was 2.1% (WORLOMETER, 2021).

Globalization, a process that is reflected by increasing the interaction between people (as well as organizations), has become one of the major characteristics of today's world (Mittelman, 1996). This social phenomenon involves many implications, e.g., commerce and especially the e-commerce market (Globerman et al., 2001), culture (Qin and Desirée, 2004), and politics (McGrew, 2005). The establishment of the European Union (EU) in 1993 might be the most significant event of globalization (Orbie and Tortell, 2009), which is the domain where the empirical evaluation of this research takes place. The reality of the EU introduced an almost unprecedented situation of completely open borders. Together with the mutual effect between globalization and transportation (Rodrigue, 2007), the number of people involved in physical contact or at least proximity increased drastically. The COVID-19 spread mechanism is based primarily on close contact between people and contaminated surfaces (WHO Q&A COVID-19, 2020). Thus, it seems that globalization has a directly negative effect on the spread of the COVID-19 pandemic. It seems that all significant worldwide pandemics since globalization emerged have shown at least one of the following characteristics: a) a non-global society, e.g., the Ebola outbreak in 2014, which started in Guinea, a country located at the bottom of the Human Development Index (HDI) (Kaner and Schaack, 2016); b) with an infection mechanism other than just proximity, e.g., AIDS which is mainly transmitted by sexual contact or exposure to bodily fluids (De Cock et al., 2012). Considering all epidemics apart from COVID-19 in the last 50 years (since 1970) excluding HIV, the average death per pandemic is about 5,500 people, with a maximum value of 284,000 in the 2009 swine flu pandemic (CDC, 2020) (LePan, n.d.). Thus, it might be that with COVID-19, humanity is experiencing for the first time a worldwide pandemic with a spread mechanism based on proximity in the new era of globalization.

[*] Corresponding author.
E-mail address: ronyh@ariel.ac.il (R.S. Hirschprung).

https://doi.org/10.1016/j.heliyon.2021.e07416
Received 24 January 2021; Received in revised form 31 March 2021; Accepted 23 June 2021

Until December 2020, there was no vaccine or efficient drugs available for COVID-19, and as of late January 2021, less than 5% of the world population had received the vaccine (OWid COVID-19, 2021; Grenfell and Drew, 2020). Thus, fighting the COVID-19 pandemic is still based primarily on prevention, such as enforcing social distancing to "flatten the curve" (Gourinchas, 2020). However, these countermeasures have a significant negative impact on social life, especially on the economy (Pike et al., 2014). An overly moderate response will reduce the economic fallout but can lead to a surge in the epidemic (Hur, 2020). In contrast, more severe lockdown measures are likely to slow the disease but harm the economy. Therefore, if not implemented carefully, the current COVID-19 counter measures have the potential of creating economic crisis and recession. Moreover, social distancing reduces the workforce and causes significant job losses (Nicola et al., 2020) (Fernandes, 2020). The COVID-19 pandemic introduces a trade-off between the values of welfare (economy) and the value of life (Li et al., 2020). A discussion on the "right" policy to balance this trade-off is a major theme (VanderWeele, 2020).

The spread of disease as a consequence of increasing globalization has been widely researched, especially the effect of motion in the modern world (Pray et al., 2006). For example, studies have explored the effect of transport networks on infection (Newman, 2002). Other studies investigated the close relationship between mortality and the number of diagnosed cases (Sarkodie and Owusu, 2020; Singh et al., 2020). Many spread models rely on the micro internal dynamics of objects such as individuals (Sahneh and Scoglio, 2011). While this theoretical approach may contribute to a better understanding of the spread mechanism, it is difficult to implement as a predictor. A more applicative approach examines the variety of factors that can affect the spread of disease (Scheidegger and Galv, 2017; Singhal et al., 2020), or by establishing a mathematical model based on these factors (Mandal et al., 2020). Even though only scant epidemiological data are publicly available for COVID-19 to date, researchers from all fields are making enormous efforts to predict infection rates, viral reproduction, death rates, and transmission patterns (Achterberg et al., 2020; Collins et al., 2020; Sperrin and McMillan, 2020). In particular, works on the transmission potential and virulence of COVID-19 (Mizumoto et al., 2020), and early prediction of outbreaks using for example the crawling of Twitter data (Jahanbin and Rahmanian, 2020; Lopez et al., 2020), have harnessed deep learning methods for automatic detection using X-ray images and Deep Convolutional Neural Networks (Alazab et al., 2020; Arora et al., 2020). Notably, these methods are standard in the field of pandemic prediction and have been successfully applied to pandemics such as the Swine flu (Ritterman et al., 2009) and H1N1 influenza (Malik et al., 2011; Lee et al., 2015). For example, Kilpatrick et al. (2006) studied the spread of H5N1 avian influenza and the phylogenetic relationship between viral isolates in birds; however, our work focuses on people. Some works rely on time-series models in an attempt to create a forecast, e.g., by adopting ARIMA (Roy et al., 2020) or cluster-based models (Ravinder et al., 2020). A study by Morse et al. (2012) gave an overview of efforts to predict pandemics, target surveillance to the most crucial interfaces, and identify prevention strategies. They showed that new mathematical modeling, diagnostics, communications, and informatics technologies could be useful in identifying and reporting hitherto unknown microbes in other species.

The COVID-19 pandemic has emerged in the digital age, characterized by an unprecedented stream of raw data that may be used for research (Gary and Nicolas, 2013; Carter and Sholler, 2016). Given the availability of both computational power and potentially useful data (epidemiologic and demographic), it seems that Data Mining methodologies are best suited to predicting the spatial spread of COVID-19 based on aggregated data as well as in other medical fields (Chen et al., 2006), and especially when parameters are tuned over time (Santosh, 2020). Unlike past pandemics, where the analysis was solely based on epidemiological data, today's understanding of the spread mechanisms can also draw on massive amounts of data from social networks, health organization updates, and news websites.

This research is based on the concept of knowledge discovery, i.e., the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996). We introduce a methodology for the dynamic prediction of the spread of diseases such as COVID-19 in a spatio-temporal space. A prediction at a specific point in time is based on the aggregated data up to this point and more general data that are typical to the region, i.e., transportation volume, national health service public investment and infrastructure. We introduce a novel indicator – the Center of Infection Mass (CoIM) – an idea inspired by gravitational physics, and use it as a parameter for simulating the future spread of the pandemic. An efficient predictor of the spread of pandemic with infection type based on close contact between people can mitigate the inherent trade-off between life-saving and economic welfare. Thus, our work can be a powerful tool in the hands of governments and decision-makers when shaping this two-pronged policy.

2. Center of infection mass (CoIM)

The COVID-19 pandemic, as mentioned above, is spread by close contact between people. Thus, we hypothesize that spread is correlated with a variety of geographic and demographic parameters (as detailed in Section 3). We also assume that the distance between a region (country in our empirical model) and a virtual center of the outbreak affects the spread of the disease. We coin this virtual center as the "Center of Infection Mass" (CoIM), an idea inspired by gravitational physics. In physics, the *center of mass* is a position defined relative to an object (or objects), which is the average position of all the parts weighted according to their masses. Given n objects in a two-dimensional space, where the i's object has mass m_i, and is located on a cartesian coordinate system with (x_i, y_i) $i \in (1..n)$, the center of mass of this system is given by: $X_{COM} = \frac{\sum_{i\in n} x_i \cdot m_i}{\sum_{i\in n} m_i}$, $Y_{COM} = \frac{\sum_{i\in n} y_i \cdot m_i}{\sum_{i\in n} m_i}$. In the same way, if we have n "objects", e.g., countries, each with confirmed COVID-19 cases c_i, and the location of each object is (x_i, y_i), we analogue the mass m_i to the confirmed cases c_i, and thus the CoIM is given by:

$$(X, Y)_{CoIM} = \left(\frac{\sum_{i\in n} x_i \cdot m_i}{\sum_{i\in n} m_i}, \frac{\sum_{i\in n} y_i \cdot m_i}{\sum_{i\in n} m_i} \right) \tag{1}$$

The idea of CoIM is demonstrated in Figure 1. In this example, the CoIM is calculated as of Apr. 2, 2020, considering: Spain, France, Belgium, Netherlands, Germany, Switzerland, Austria, and Italy. In the left figure (a), each yellow circle indicates the proportion of confirmed cases (c_i) relative to the total amount of confirmed cases in the countries considered $(\sum_{\forall j} c_j)$. Thus, each circle surface is: $S_i \propto c_i \cdot \sum_{\forall j} c_j$. The red crossed marker indicates the location of the CoIM. The movement of the CoIM across time is depicted in Figure 1(b).

We hypothesize that accommodating the CoIM index in the model may increase the prediction accuracy mainly because of the means of transferring the pathogen (the organism that causes the disease). COVID-19 infection requires proximity between people since its main infection mechanism is via exposure to droplets (Odor et al., 2020). The infection is believed to occur by exposure to symptomatic patients, as well as pre-symptomatic and asymptomatic patients. All three are diagnosed by the COVID-19 tests (although there are some false-negatives, especially in the early stages of the disease) and expressed in the confirmed cases index. Thus, the infection process is stochastic by its nature, i.e., when a person is exposed to a population that contains carriers of the disease, his infection probability is positively correlated with the rate of carriers surrounding him. This factor can be evaluated through the CoIM index, which describes the infection rate around a specific point. Moreover, the effect of the CoIM index on the prediction accuracy is expected to increase when the resolution of the data is increased (as described in section 5). The rationale behind this claim is that when the resolution is increased, i.e., smaller cells of data, the CoIM better indicates the rate of carriers around a specific point, thus, better indicates the probability of being infected. This probability is directly pursuant to the number of infected people, which is the prediction target.

3. Prediction methodology

The methodology we offer aims to predict the number of confirmed cases in each region based on past data. In practice, the predicted variable was chosen to be the growth factor. The growth factor is defined as the factor by which the number of confirmed cases multiplies itself over a time period, and is $c_i^{t+\Delta}/c_i^t$ when c_i^t is the number of confirmed cases of country i at time t, and Δ is the time intervals, for example, one day or one week. The prediction process is divided into four major phases: a) data collection; b) preprocessing; c) training; d) actual prediction, as depicted in Figure 2.

In the first phase, data is collected from public sources, e.g., GIS and Maps at Johns Hopkins University (Hopkins Libraries, 2020). Usually,

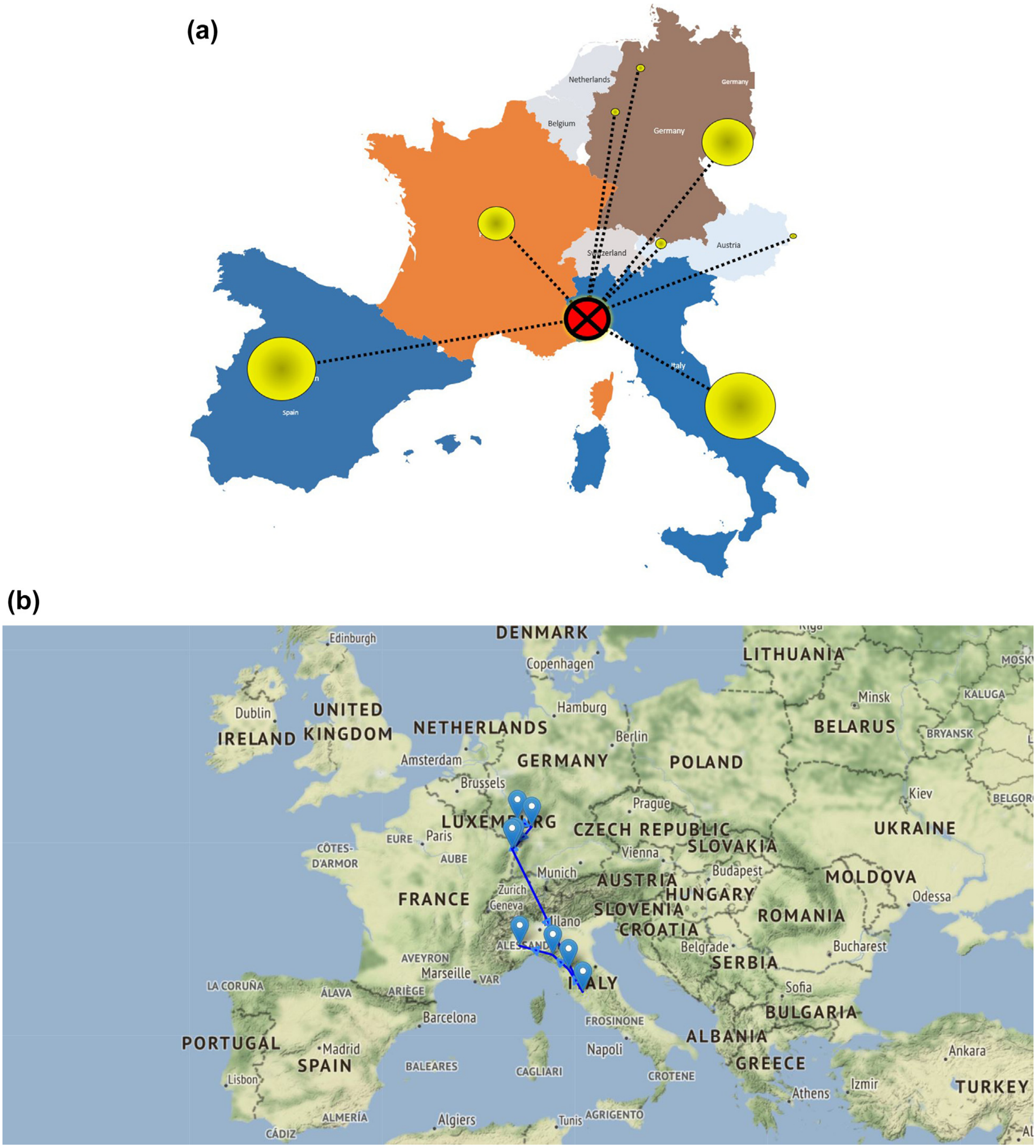


Figure 1. Demonstrating the idea of the CoIM (Center of Infection Mass) index. (a) The left figure provides static information as of Apr. 2, 2020. Each yellow circle indicates the proportion of confirmed cases relative to the total amount of confirmed cases in the countries considered. The red crossed marker indicates the location of the CoIM. The lines visualize how each country draws the CoIM towards itself; (b) The right figure demonstrates the movement of the CoIM across time. The period described in the figure is between Feb. 2, 2020 to Apr. 26, 2020 with time intervals of 1 week. The yellow arrows indicate the movement direction. The small map in the top right corner is displayed for orientation.

data is available on a daily basis; however, this relatively high resolution may introduce local peaks into the model, e.g., cases drop with fewer tests over the weekend (IANS, 2020). Thus, we adopted two methods that are applied in the data preprocessing phase. The first one is *averaging* the data over a time period, e.g., a week, and for country i is: $\sum_{r\in t} c_i^r / |t|$. The second one is *exponential smoothing* of the data over the time period according to the equation:

$$\begin{cases} \textit{not counted} & t<I \\ \left[\sum_{r=2}^{I}\alpha(1-\alpha)^{I-r}c_i^r\right]+(1-\alpha)c_i^1 & t=I \\ \alpha c_i^t+(1-\alpha)c_i^{t-1} & t>I \end{cases} \tag{2}$$

where I is an initiation period of observations that are not included in the training model but considered when calculating observations beyond this period; and α is the smoothing factor ($0.5<\alpha<1$). The larger the α factor is, the less smoothing is applied.

The data-independent variables that are used for the prediction model may be classified into two groups. The first one contains static data – variables which are not changing across time – and includes: road passenger transport by buses and coaches (normalized), rail passenger transport (normalized), flights transportation (yearly), population density (ppl/km2), HDI index, social spending (% of GPD), GDP per capita, border countries, health expenditure (%of GPD), share of 65-year-olds and above, share of under 15-year-olds, share of 15 to 24-year-olds, share of 25 to 64-year-olds, medical doctors per 10,000 ppl, and medical beds per 10,000 ppl. The second group of variables contains dynamic data – variables that may change across time – and includes: weekly distance from the CoIM, weekly confirmed cases, weekly deaths, and weekly tests done. Some of the features are hypothesized to indicate higher confirmed cases: a) flights transportation which indicates people moving from country to country; b) population density that may increase the rate of infection through a higher rate of contacts; and c) share of the 65-year-olds in the population which relates to a population which is more susceptible to infection. On the other hand, countries characterized by a higher rate of medical doctors and a higher rate of medical beds (measured per 10,000 ppl) are assumed to have a better chance of providing good medical care and higher awareness of healthcare, resulting in fewer confirmed cases. It is worth mentioning that while the causality of some factors may be explained or at least assumed, many others are still studied by researchers. For example, it was found that the rate of exhaled aerosol, which is the means through which COVID-19 is spread, is highly correlated with age but not with sex (Edwards et al., 2021). Since our model does not require establishing the causality, it introduces a significant advantage at the early stages of an outbreak, when very little epidemiologic information is available.

In the training phase, we seek to discover the underlying knowledge of predicting the number of confirmed cases at time period $t+1$ based on the data from time period t and before. We applied different machine learning models to the preprocessed data to yield rules that establish the required 1 time-step prediction. We evaluated a wide range of prediction models, including linear regression, Random Forest Regressor (RF), XGBoost Regressor, Ridge regression, Lasso regression, and Support Vector Regressor (SVR). The Ridge and Lasso models were evaluated using two possible values for α , 1 and 50.

In the actual prediction phase, we simulate the process a few steps forward. Given data till time period t, we predict the confirmed cases for time period $t+1$, and then, based on the aggregated data including that of $t+1$ we predict the confirmed cases for $t+2$ and back again. The prediction simulations were conducted using two approaches: a) *Timeline data only*: In this approach, we trained the data for each country isolated from the other countries; however, the interaction between countries is considered through the CoIM index; and b) *Timeline & static data*: In this approach, we first "flattened" the database so that each record includes: time, country, static data, timeline data (where time and country construct the primary key). Then, we trained the data for the whole country set. The *Timeline data only* approach is expected to yield more accurate results, while the *Timeline & static data* approach has the potential of predicting a new country that was not introduced to the original model. To evaluate the error of the prediction, we used a defined index named ACC, which averages the errors across all time predictions:

$$ACC=\frac{1}{N}\sum_{\forall t}\frac{\left|ca_i^t-cp_i^t\right|}{ca_i^t} \tag{3}$$

where ca_i^t and cp_i^t are the actual confirmed cases and predicted confirmed cases for country i at time period t respectively. Naturally, as one can expect, long-term predictions (e.g., five weeks vs. two weeks) result in a higher deviation from the true reality.

4. Empirical evaluation

To empirically evaluate the model, we selected the western European region as the experiment environment. Countries in this region are characterized by open borders, and data is considered reliable. The COVID-19 pandemic hit hard in Europe; for example, by late October 2020, Spain passed 1M confirmed cases (Jones, 2020). The countries we included in the empirical evaluation are France, Spain, Italy, Germany, Austria, Netherlands, Belgium, Greece, Finland, Switzerland.

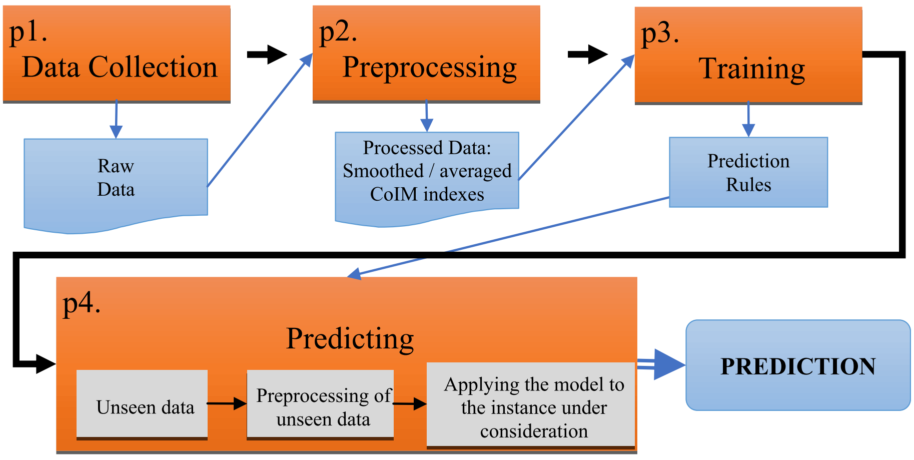


Figure 2. This figure describes the four major phases of the prediction process (in orange boxes), and their products (in blue boxes).

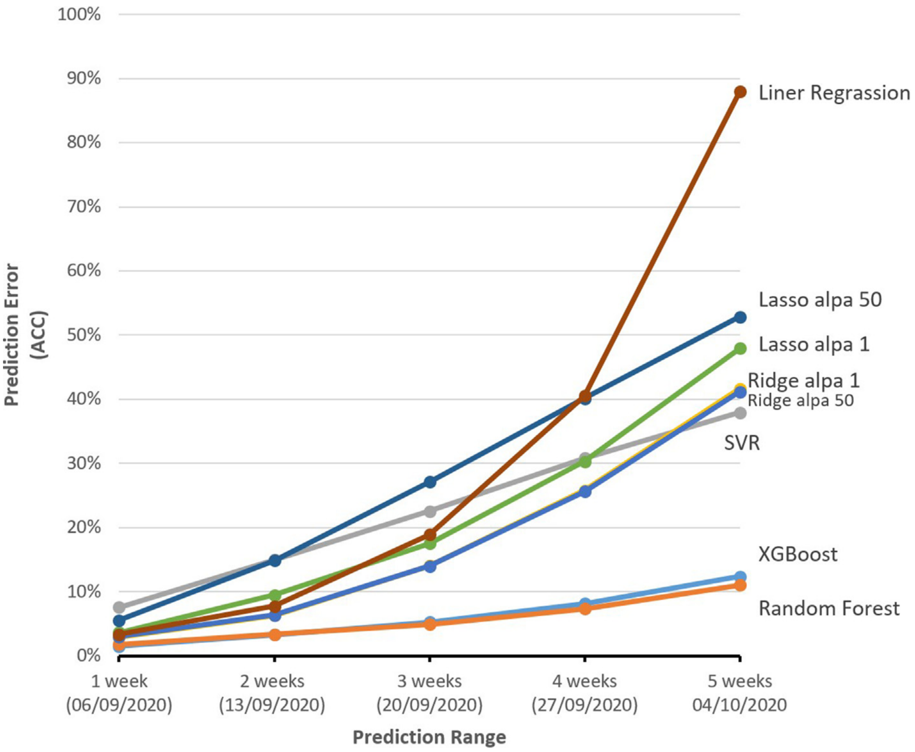


Figure 3. This figure describes the prediction accuracy (by the error) of the various models. The X-axis describes the prediction range, while the Y-axis describes the prediction error (ACC). The higher the error, the lower the accuracy.

Models and simulations were developed using Python 3 on Jupyter Notebook environment and Spyder integrated development environment for math programming. The machine learning code was implemented using the Sci-kit learn Python library and XGBoost library[1]. Microsoft Excel was used for data preprocessing, and map production was executed using Microsoft 3D Map and Folium library. Dynamic data on the number of confirmed cases, deaths, and number of tests done, were collected automatically from Johns Hopkins University (Hopkins Libraries, 2020), by using Python procedures available on their GitHub (Hopkins GitHub, 2021). On the other hand, static data was manually crawled: general static medical data from WHO (WHO-beds, 2020) (WHO-doctors, 2020), static population data from The World Bank (TWB-population, 2020) (TWB-density, 2020), economic data from the World Bank (TWB-economy, 2020), health spending from the OECD (OECD-data, 2020), and passenger air traffic from Wikipedia (Wikipedia-trafic, 2020) and from World by Map (WbM, 2020). The training data was composed of data from Apr. 5, 2020 to Aug. 30, 2020, and the prediction test period was Sep. 6, 2020 to Oct. 4, 2020, resulting in a prediction range of five weeks. We define 1–2 weeks as *short range*, 3 weeks as *medium range*, and 4–5 weeks as *long range*. We also define three discrete levels of prediction: *High accuracy* when $ACC \leq 5\%$, *Mid accuracy* when $5\% < ACC \leq 10\%$), and *Low accuracy* when $10\% < ACC$.

Initially, we applied a simple linear regression that achieved a *high* to *mid accuracy* prediction for short range ($3\% \leq ACC \leq 8\%$), but *low accuracy* prediction for the medium and long range ($19\% \leq ACC \leq 88\%$). We then applied advanced algorithms, including: Random Forest, XGBoost, Ridge ($\alpha = 1$), Ridge ($\alpha = 50$), Lasso ($\alpha = 1$), and Lasso ($\alpha = 50$). Random Forest consists of training multiple regression decision trees, each with access to a different subset of the data, and use averaging of different prediction decision trees to improve the predictive accuracy and the control of overfitting. XGBoost is another regression model which is based on decision trees. While Random Forest can be considered as a parallel model based on the wisdom of the crowd (i.e., the average prediction among different trees), XGBoost adapts a more sequential approach, when each decision tree is aimed at learning from the errors of its predecessor. In Ridge regression, we add a penalty term that is equal to the square of the coefficient (i.e., feature importance and effect on the number of confirmed cases). Ridge regression decreases the complexity of a model. Still, it does not reduce the number of variables since it never leads to a coefficient of zero but only minimizes it. Therefore, we also included the Lasso regression models that add a penalty term onto the absolute sum of the coefficients. The difference between Ridge and Lasso regression is that the latter tends to make coefficients absolute zero compared to Ridge, which never sets the value of coefficients to absolute zero. For both Ridge and Lasso regression, the parameter α controls the trade-off between low error and low coefficients value. The prediction accuracy of the various algorithms is depicted in Figure 3. As mentioned above, the Linear Regression, which is the most simplistic model, yields relatively poor results. In contrast, more sophisticated models such as XGBoost and Random Forest yield the best results.

Data items of this type tend to include local noise; thus, we applied smoothing by both the average smoothing and the exponential smoothing methods. A comparison of the prediction accuracy of both methods is depicted in Figure 4. The figure shows that the average mean smoothing method has superiority over the exponential smoothing methods in the short and medium prediction range. However, the exponential smoothing method is superior in the long prediction range. In any case, the differences are not highly significant, they might be an outcome of a random noise, and there is not enough evidence to make a determination regarding the preferred smoothing method.

We then applied the Random Forest algorithm to the dataset mentioned above, with the average smoothing method. The accuracy of the prediction according to the country predicted and the prediction range is depicted in Figure 5, and the average prediction error across all countries is depicted in Figure 6. It can be noticed that for the short and medium ranges, all countries have a high accuracy prediction apart from

[1] The prediction codes used in this research are publicly available at https://github.com/chenhajaj/COVID19PredictionModel, for reusability purposes.

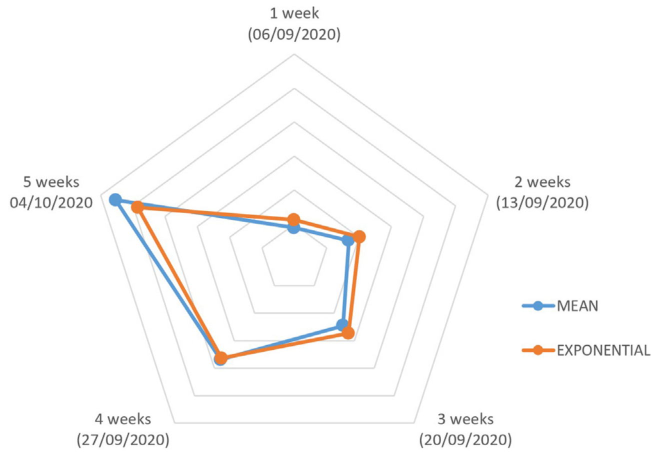


Figure 4. Exponential vs. the Average Smoothing Methods. This figure compares the exponential and the average smoothing methods. The orange polygon indicates the prediction error (ACC) with exponential smoothing, while the blue polygon indicates mean smoothing for different prediction ranges. The distance of each point from the chart center indicates the error level.

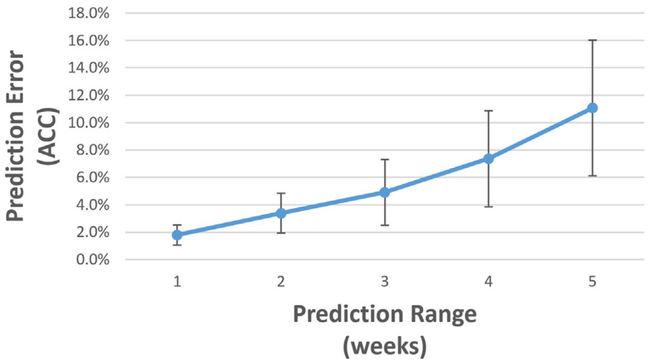
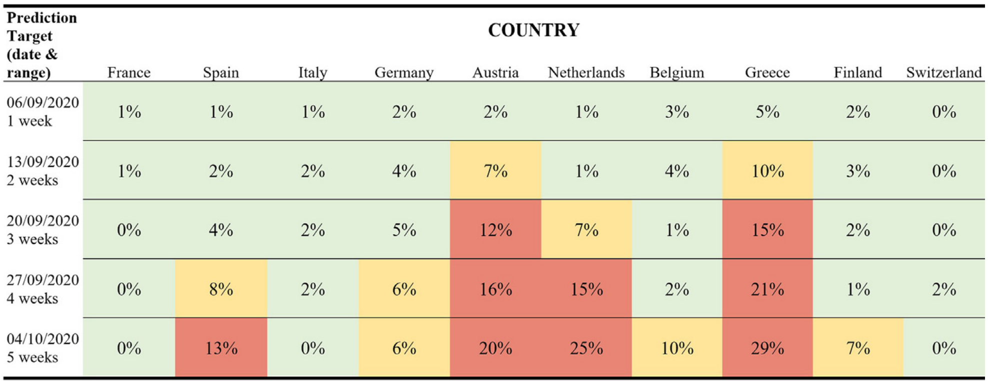


Figure 6. This figure describes the average prediction error across all countries. The X-axis describes the prediction range (in weeks), while the Y-axis describes the prediction error (ACC). The whiskers describe the standard deviation of ACC across different countries.

Table 1. The two most important features for each model.

COUNTRY									
France	Spain	Italy	Germany	Austria	Netherlands	Belgium	Greece	Finland	Switzerland
Weekly confirmed	Weekly distance	Weekly confirmed	Weekly confirmed	Weekly distance	Weekly confirmed	Weekly deaths	Weekly distance	Weekly confirmed	Weekly deaths
Weekly deaths	Weekly confirmed	Weekly deaths	Weekly deaths	Weekly test done	Weekly deaths	Weekly confirmed	Weekly deaths	Weekly deaths	Weekly test done



Prediction Target (date & range)	COUNTRY									
	France	Spain	Italy	Germany	Austria	Netherlands	Belgium	Greece	Finland	Switzerland
06/09/2020 1 week	1%	1%	1%	2%	2%	1%	3%	5%	2%	0%
13/09/2020 2 weeks	1%	2%	2%	4%	7%	1%	4%	10%	3%	0%
20/09/2020 3 weeks	0%	4%	2%	5%	12%	7%	1%	15%	2%	0%
27/09/2020 4 weeks	0%	8%	2%	6%	16%	15%	2%	21%	1%	2%
04/10/2020 5 weeks	0%	13%	0%	6%	20%	25%	10%	29%	7%	0%

Figure 5. Prediction Error of the Selected Model. This figure introduces the prediction accuracy by the error (ACC) yield by Random Forest algorithm, for timeline data only approach and average smoothing method. For each prediction target date (indicated by the actual date and the prediction range in weeks), and for each country, the ACC error index is reported. The higher the error, the lower the accuracy. Green cells indicate high accuracy, yellow cells mid accuracy, and red cells low accuracy.

Austria and Greece, which have a mid to low accuracy prediction. As this is not the focus of this study, we leave the question of why these countries behave differently for future research. For the long range, half of the countries have a high accuracy prediction, and the other half have a mid to low accuracy prediction.

Analyzing the feature importance of the various variables, it can be noted that for the models of Germany, Italy, Belgium, Finland, and Netherlands, the two most important features were the number of weekly confirmed cases, and weekly number of deaths. Interestingly, Spain, Austria, and Greece all share the weekly distance from the CoIM as one of their two most important features. The accumulated results of the two most significant features for each country are depicted in Table 1.

Notably, and as expected, the longer the prediction range, the lower the prediction accuracy.[2] This phenomenon is also explained by the deviation of the predicted CoIM from the actual CoIM as depicted in Figure 7. Our next evaluation shows the importance of the CoIM index in the models. This time, we trained a Random Forest model for each country, but excluded the weekly distance feature (i.e., ignored the CoIM).

As depicted in Figure 8, for a prediction range of one week, there is an insignificant advantage to the model without the CoIM index, but for a prediction range of two weeks and above, the model that takes the CoIM index into consideration is much more accurate.

Finally, we applied the *timeline & static data* approach by "flattening" the DB as explained in the Prediction Methodology section. As depicted in Figure 9, it can be seen that the prediction accuracy of the model based on *timeline data only* is significantly higher than the model based on *timeline & static data*. Interestingly, similar to the country-based models,

[2] Considering this regularity, France shows a slight anomaly with lower accuracy for the short range and higher accuracy for the mid and long range – however, the deviation is minor.

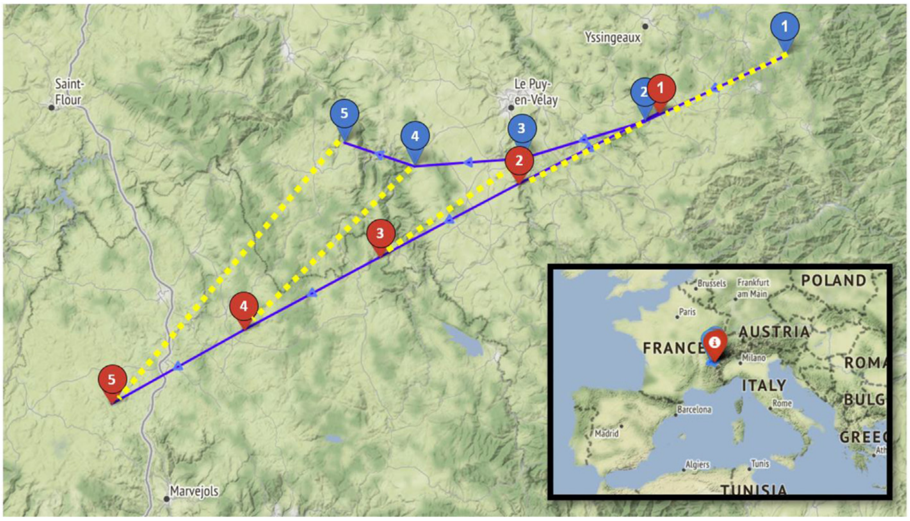


Figure 7. This figure describes the movement of the predicted CoIM vs. the movement of the actual CoIM. The red markers indicate the prediction and are numbered ordinally according to the time range (weeks), while the blue markers indicate the reality and are numbered in the same way. The broken yellow lines stretched between two paired markers (predicted vs. reality) visualize the distance. The small map in the bottom right corner is displayed for orientation.

in this model as well, the two most important features were weekly number of confirmed cases and weekly number of deaths.

5. Discussion

This research introduces a methodology to predict the spread of the COVID-19 pandemic in a spatial-temporal space. The prediction model is based on a novel concept – The Center of Infection Mass (CoIM), inspired from the physics center of mass – when mass centers at a given location are replaced with infection extents in their 2D geographic coordinates. We trained various regression models, e.g., Random Forest, XGBoost with historical data of a specific period, and predicted the progress of the infection in a followed period. Our results show that for the range of one to three weeks, most countries' COIVID-19 spread could be predicted with high accuracy, and for a range of four to five weeks with lower accuracy.

The ability to predict the spread of disease may mitigate the inherent trade-off between life-saving and economy; both are fundamental social values (Li et al., 2020), and may also influence policy economic design, e.g., providing stimulus packages (Siddik, 2020), which may contribute significantly to society's welfare. Many, if not most of the pandemics, similar to the case of COVID-19, are unexpected (Sturmberg and Martin, 2020). Therefore, in the gap between the time when knowledge of the existence of the pandemic is discovered and when an effective vaccine of medicine is found, applying a balanced policy may be the only tool in the hands of decision-makers.

The analysis in this paper is based on publicly published results, which are highly available in the COVID-19 pandemic outbreak event (Lin and Hou, 2020). However, government officials may collect and have access to more detailed data, e.g. by tracking mobile phones (Calvo et al., 2020). Moreover, data can be collected based on self-disclosing, e.g., text-mining from unusual messages (Ku et al., 2014). Notably, this

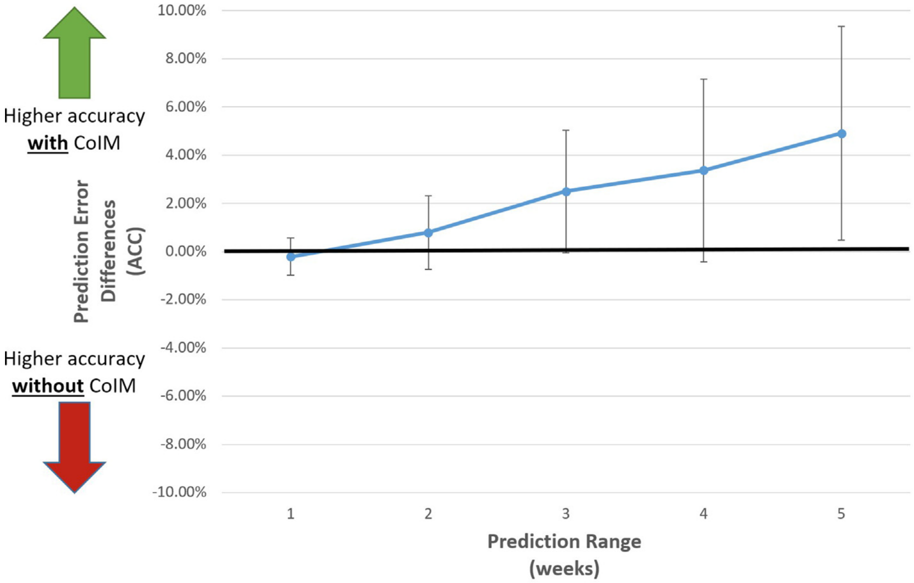


Figure 8. This figure describes the differences between the prediction with and without the CoIM index. The X-axis describes the prediction range in weeks, while the Y-axis describes the difference in the prediction error (ACC). A positive value of the difference indicates a better prediction by the model with the CoIM index, and a negative value the opposite.

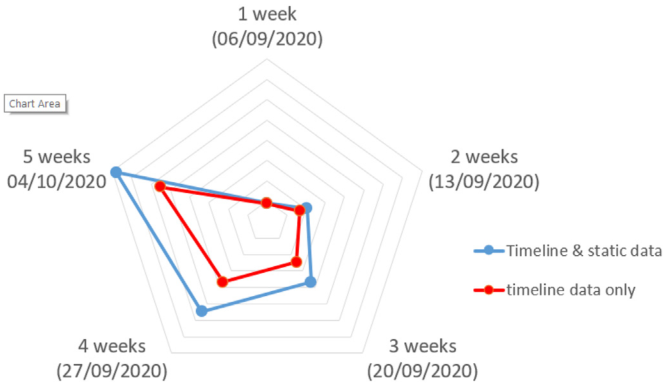


Figure 9. This figure compares the timeline data only vs. timeline and static data approaches. The red polygon indicates the prediction error (ACC) of the model based on the timeline data only approach, while the blue polygon indicates the prediction error of the model based on the timeline & static data approach. The distance of each point from the chart center indicates the error level.

act may violate privacy (Thompson et al., 2020), and it is in the center of public attention in democratic countries (Ienca and Vayena, 2020), and introduces another trade-off – here vs. the value of privacy. The availability of the data in a higher resolution, i.e., in the spatio-temporal space with lower granularity (e.g., street-level instead of country level and/or hourly time intervals instead of daily intervals) may enable the application of our prediction model at a micro-level. An outcome of this knowledge can be, for example, posing restrictions in relatively small regions, thus minimizing the social and economic hit. This strategy is aligned with the idea of a “traffic light” road map which was adopted by many countries, e.g. the EU (Riegert, 2020). Applying our model with higher resolution data is straightforward; however, further research is required to validate the prediction accuracy.

The model was applied in a mature state, when the virus had already spread, and all countries in the region were infected. However, the *timeline & static data* can theoretically be applied at a preliminary state of the pandemic spread because the prediction may be based on the static data which characterized a country or region (available before the disease hit) and the CoIM is derived from neighboring countries which were already infected. The use of this model may be further researched. This availability is handy in fighting pandemics at their early stages and maybe even avoiding widespread contagion.

The model is suitable for any pandemic with a spread mechanism based on proximity. However, the prediction may also be tested on other types of pandemics, such as zoonotic disease (diseases that pass from animals or insects to a human) like Leptospirosis, a disease with a global effect (Bharti et al., 2003). This may be achieved by accommodating other relevant parameters; however, in some cases, e.g., genetic disease, the CoIM concept may not be relevant. For future research of the COVID-19 pandemic as well as others, we highly encourage the development of fully-automated data collection mechanisms with a friendly user-interface, that might be of great importance for policymakers and fellow researchers.

This research has two limitations: The first one is that the model assumes continuity and does not consider events like blocking air traffic. Events of this type introduce a point of discontinuity, which may reduce model accuracy. This limitation may be addressed in the following way: If the discontinuity point belongs to the past, it is better to train the model from this point on; however, if the discontinuity point occurs in the predicted period, further research is required to model its effect. The second limitation is that the model relies on the reliability of the data, and as mentioned above, data-collection itself is at the heart of public debate due to the potential of privacy violation and limits of government power.

The prediction methodology we introduce in this paper, which was tested on the COVID-19 pandemic, establishes the foundations for: further research to enhance this model; increase its accuracy and prediction range and to apply it in other cases; and to build a machine learning-based application that will be applied in real-time, and will contribute to the improvement of human welfare during pandemic outbreaks.

Declarations

Author contribution statement

R. S. Hirschprung, C. Hajaj: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by Ariel University (IL) grant no. RA2000000303.

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgement

This research was supported by Ariel University, COVID-19 call for research, grant number RA2000000303. We thank the Ariel Cyber Innovation Center in conjunction with the Israel National Cyber directorate in the Prime Minister’s Office. We would also like to thank Mr. Eyal Yahav for his assistance with data collection programming and analysis.

References

Achterberg, M. A. a, P., B., et al., 2020. Comparing the accuracy of several network-based COVID-19 prediction algorithms. Int. J. Forecast.

Alazab, M. a, A, A., et al., 2020. COVID-19 prediction and detection using deep learning. Int. J. Comp. Info. Syst. Indust. Manag. Appl. 12, 168–181.

Arora, P., Kumar, H., Panigrahi, B.K., 2020. Prediction and analysis of COVID-19 positive cases using deep learning models: a descriptive case study of India. Chaos, Solit. Fractals 139, 110017.

Bharti, A.R., et al., 2003. Leptospirosis: a zoonotic disease of global importance. Lancet Infect. Dis. 3 (12).

Calvo, R.A., Deterding, S., Ryan, R.M., 2020. Health surveillance during COVID-19 pandemic. British Med. J. Publish. Group.

Carter, D., Sholler, D., 2016. Data science on the ground: hype, criticism, and everyday work. J. Assoc. Inform. Sci. Technol. 67 (10), 2309–2319.

CDC, 2020. Past Pandemics [Online] Available at: https://www.cdc.gov/flu/pandemic-resources/basics/past-pandemics.html.

Chen, H., Fuller, S.S., Friedman, C., Hersh, W., 2006. Knowledge management, data mining, and text mining in medical informatics. In: Medical Informatics: Knowledge Management and Data Mining in Biomedicine. s.L. Springer Science & Business Media.

CNBC, 2020. WHO Says 2 Million Coronavirus Deaths Is ‘not Impossible’ as World Approaches 1 Million [Online] Available at: https://www.cnbc.com/2020/09/25/who-says-2-million-coronavirus-deaths-is-not-impossible-as-world-approaches-1-million-.html.

Collins, G.S., van Smeden, M., Riley, R.D., 2020. COVID-19 prediction models should adhere to methodological and reporting standards. Eur. Respir. J. 56 (3).

De Cock, K.M., Jaffe, H.W., Curran, J.W., 2012. The evolving epidemiology of HIV/AIDS. AIDS 26 (10), 1205–1213.

Edwards, D.A., et al., 2021. Exhaled Aerosol Increases with COVID-19 Infection, Age, and Obesity. s.L. National Academy of Sciences.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery: an overview. Adv. Knowled. Discov. Data Mining 1 (35), 1–36.

Fernandes, N., 2020. Economic effects of coronavirus outbreak (COVID-19) on the world economy. Available at SSRN 3557504.

Gary, H., Nicolas, F., 2013. Online information: access, search and exchange. In: The Sage Handbook of Digital Technology Research. s.L. Sage.

Globerman, S., Roehl, T.W., Standifird, S., 2001. Globalization and electronic commerce: inferences from retail brokering, 32 (4), 749–768.

Gourinchas, P.-O., 2020. Flattening the Pandemic and Recession Curves. Mitigating the COVID Economic Crisis: Act Fast and Do Whatever, p. 31.

Grenfell, R., Drew, T., 2020. Here's why it's taking so long to develop a vaccine for the new coronavirus [Online] Available at: https://www.sciencealert.com/who-says-a-coronavirus-vaccine-is-18-months-away.

Hopkins GitHub, J., 2021. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [Online] Available at: https://github.com/CSSEGISandData/COVID-19.

Hopkins Libraries, S.L.J., 2020. GIS and Maps [Online] Available at: https://github.com/CSSEGISandData/COVID-19.

Hur, S., 2020. The Distributional Effects of Covid-19 and Mitigation Policies, 400. Globalization and Monetary Policy Institute Working Paper.

IANS, 2020. Telangana coronavirus update: cases drop with fewer tests over the weekend [Online] Available at: https://www.business-standard.com/article/current-affairs/telangana-coronavirus-update-cases-drop-with-fewer-tests-over-the-weekend-120092801567_1.html.

Ienca, M., Vayena, E., 2020. On the responsible use of digital data to tackle the COVID-19 pandemic. Nat. Med. 26 (4), 463–464.

Jahanbin, K., Rahmanian, V., 2020. Using Twitter and Web News Mining to Predict COVID-19 Outbreak.

Jones, S., 2020. Spain is first western European country to pass 1m Covid cases [Online] Available at: https://www.theguardian.com/world/2020/oct/21/spain-first-western-european-country-to-pass-1-million-covid-cases.

Kaner, J., Schaack, S., 2016. Understanding Ebola: the 2014 epidemic. Glob. Health 12 (1), 1–7.

Kilpatrick, A., et al., 2006. Predicting the global spread of H5N1 avian influenza. Proc. Natl. Acad. Sci. Unit. States Am. 103 (51), 19368–19373.

Ku, Y., et al., 2014. Text mining self-disclosing health information for public health service. J. Assoc. Inform. Sci. Technol. 65 (5), 928–947.

Lee, A.J., et al., 2015. Diversifying selection analysis predicts antigenic evolution of 2009 pandemic H1N1 influenza A virus in humans. J. Virol. 89 (10), 5427–5440.

LePan, N., n.d.. Visualizing the history of pandemics [Online] Available at: https://www.visualcapitalist.com/history-of-pandemics-deadliest/.

Li, H.-L., Jecker, N.S., Chung, R.Y.-N., 2020. Reopening economies during the COVID-19 pandemic: reasoning about value tradeoffs. Am. J. Bioeth. 20 (7), 136–138.

Lin, L., Hou, Z., 2020. Combat COVID-19 with artificial intelligence and big data. J. Trav. Med. 5, 27.

Lopez, C.E., Vasu, M., Gallemore, C., 2020. Understanding the Perception of COVID-19 Policies by Mining a Multilanguage Twitter Dataset.

Malik, M.T., et al., 2011. "Google flu trends" and emergency department triage data predicted the 2009 pandemic H1N1 waves in Manitoba. Can. J. Public Health 102 (4), 294–297.

Mandal, M.a.J.S., et al., 2020. A Model Based Study on the Dynamics of COVID-19: Prediction and Control. Chaos, Solitons & Fractals, p. 109889.

McGrew, A., 2005. Globalization and global politics. Globaliz. World Polit. 3, 19–40.

Mittelman, J.H., 1996. Globalization: Critical Reflections. s.L. Lynne Rienner Boulder, CO.

Mizumoto, K., Kagaya, K., Chowell, G., 2020. Early Epidemiological Assessment of the Transmission Potential and Virulence of Coronavirus Disease 2019 (COVID-19) in Wuhan City: China. medRxiv.

Morse, S.S., et al., 2012. Prediction and prevention of the next pandemic zoonosis. Lancet 380 (9857), 1956–1965.

Newman, M.E., 2002. Spread of epidemic disease on networks. Phys. Rev. 66 (1).

Nicola, M., et al., 2020. The socio-economic implications of the coronavirus pandemic (COVID-19): a review. Int. J. Jurgery (London, England) 78, 185.

Odor, P.M., et al., 2020. Anaesthesia and COVID-19: infection control. Br. J. Anaesth.

OECD-data, 2020. Health Spending [Online] Available at: https://data.oecd.org/healthres/health-spending.htm.

Orbie, J., Tortell, L., 2009. From the social clause to the social dimension of globalization. In: The European Union and the Social Dimension of Globalization: How the EU Influences the World. s.L. Routledge.

OWid COVID-19, O.W.i.D., 2021. Coronavirus (COVID-19) Vaccinations [Online] Available at: https://ourworldindata.org/covid-vaccinations.

Pike, J., et al., 2014. Economic optimization of a global strategy to address the pandemic threat. Proc. Natl. Acad. Sci. Unit. States Am. 111 (52), 18519–18523.

Pray, L., et al., 2006. A world in motion: the global movement of people, products, pathogens, and power. In: The Impact of Globalization on Infectious Disease Emergence and Control: Exploring the Consequences and Opportunities: Workshop Summary. s.L. National Academies Press.

Qin, M.M., Desirée, S.-O., 2004. Imperial feelings: youth, culture, citizenship, and globalization. In: Globalization: Culture and Education in the New Millennium. s.L. Univ of California Press.

Ravinder, R., et al., 2020. An Adaptive, Interacting, Cluster-Based Model for Predicting the Transmission Dynamics of COVID-19. Heliyon, e05722.

Riegert, B., 2020. Coronavirus: what the EU's New Traffic Light System Means [Online] Available at: https://www.dw.com/en/coronavirus-what-the-eus-new-traffic-light-system-means/a-55265476.

Ritterman, J., Osborne, M., Klein, E., 2009. Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic, 9. 1st International Workshop on Mining Social Media, pp. 9–17.

Rodrigue, J.-P., 2007. Transportation and Globalization. Encyclopedia of Globalization.

Roy, S., Bhunia, G.S., Shit, P.K., 2020. Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. Model. Earth Syst. Environ. 1–7.

Sahneh, F.D., Scoglio, C., 2011. Epidemic Spread in Human Networks. s.L. IEEE, pp. 3008–3013.

Santosh, K., 2020. COVID-19 prediction models and unexploited data. J. Med. Syst. 44 (9), 1–4.

Sarkodie, S.A., Owusu, P.A., 2020. Investigating the Cases of Novel Coronavirus Disease (COVID-19) in China Using Dynamic Statistical Techniques. Heliyon, e03747.

Scheidegger, A., Galv, P., 2017. A Mathematical Approach for Classifying Input Parameters for Infectious Disease Spread. s.l. Institute of Industrial and Systems Engineers (IISE), pp. 1060–1065.

Siddik, M.N.A., 2020. Economic stimulus for COVID-19 pandemic and its determinants: evidence from cross-country analysis. Heliyon 6 (12), e05634.

Singhal, A., Singh, P., Lall, B., Joshi, S.D., 2020. Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. Chaos, Solit. Fractals 138, 110023.

Singh, V., et al., 2020. Prediction of COVID-19 corona virus pandemic based on time series data using Support Vector Machine. J. Discrete Math. Sci. Cryptogr. 1–15.

Sperrin, M., McMillan, B., 2020. Prediction Models for Covid-19 Outcomes. s.L. British Medical Journal Publishing Group.

Sturmberg, J.P., Martin, C.M., 2020. COVID-19–how a pandemic reveals that everything is connected to everything else. J. Eval. Clin. Pract.

Thompson, N., McGill, T., Bunn, A., Alexander, R., 2020. Cultural factors and the role of privacy concerns in acceptance of government surveillance. J. Assoc. Inform. Sci. Technol. 71 (9), 1129–1142.

TWB-density, 2020. Population Density (People Per Sq. Km of Land Area) [Online] Available at: https://data.worldbank.org/indicator/EN.POP.DNST?name_descfalse&viewmap.

TWB-economy, 2020. Countries and Economies [Online] Available at: https://data.worldbank.org/country.

TWB-population, 2020. Population, Total [Online] Available at: https://data.worldbank.org/indicator/SP.POP.TOTL.

VanderWeele, T.J., 2020. Challenges estimating total lives lost in COVID-19 Decisions: consideration of mortality related to unemployment, social isolation, and depression. J. Am. Med. Assoc. 324, 445–446.

WbM, 2020. Passenger SSENGER TRAFFIC BY AIR [Online] Available at: https://www.citypopulation.de/en/world/bymap/AirTrafficPassengers.html.

WHO Q&A COVID-19, W.H.O., 2020. Q&A on Coronaviruses (COVID-19) [Online] Available at: https://www.who.int/news-room/q-a-detail/q-a-coronaviruses.

WHO-beds, 2020. THE GLOBAL HEALTH OBSERVATORY - Beds Per 10,000 [Online] Available at: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/hospital-beds-(per-10-000-population.

WHO-doctors, 2020. THE GLOBAL HEALTH OBSERVATORY - Medical Doctors Per 10,000 [Online] Available at: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/medical-doctors-(per-10-000-population.

WHO, W.H.O and others, 2020. Coronavirus Disease 2019 (COVID-19): Situation Report, 59, s.L. World Health Organization.

Wikipedia-trafic, 2020. List of Countries by Airline Passengers [Online] Available at: https://en.wikipedia.org/wiki/List_of_countries_by_airline_passengers.

WORLOMETER, 2021. COVID-19 CORONAVIRUS PANDEMIC [Online] Available at: https://www.worldometers.info/coronavirus/.

Zu, Z.Y., et al., 2020. Coronavirus disease 2019 (COVID-19): a perspective from China. Radiology.