Susceptible, Infectious and Recovered (SIR Model) Predictive Model to Understand the Key Factors of COVID-19 Transmission

DeepaRani Gopagoni[1], P V Lakshmi[2]
Computer Science and Engineering
GITAM Institute of Technology
Vishakhapatnam, Andhra Pradesh
India

***Abstract*—On 31 December 2019, WHO was alerted to several cases of pneumonia in Wuhan City, Hubei Province of China. The virus did not match any other known virus. This raised concern because when a virus is new, general behavior and how it affects, people do not know. Initial few cases reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly have not had exposure to animal markets, indicating person-to-person spread is occurring. At this time, it's unclear how easily or sustainably this virus is spreading between people. At any given time during a flu epidemic, firstly, should know the number of people who are infected. Second, to know the numbers who have been infected and have recovered, because these people now have immunity to the disease. Well established SIR modeling methodology is used to develop a predictive model in order to understand the key factors that impact the COVID-19 transmission.**

***Keywords*—COVID-19; SIR modeling; WHO; disease spread**

I. Introduction

In December 2019 World Health Organization alerted to several cases of pneumonia in Wuhan City, Hubei Province of China [1]. The virus did not match any other known virus. Novel Corona virus (2019-nCoV) is a virus (more specifically, a corona virus) identified as the cause of an outbreak of respiratory illness. Early on, many of the patients in the outbreak in Wuhan, China reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly have not had exposure to animal markets, indicating person-to-person spread is occurring [2-8]. This raised concern because when a virus is new, does not know how it affects people. At this time, it's unclear how easily or sustainably this virus is spreading between people [9]. Moreover, according to one study, presumed hospital-related transmission of SARS-CoV-2 was suspected in 41% of patients [8]. Based on the evidence of a rapidly increasing incidence of infections [11] and the possibility of transmission by asymptomatic carriers [12], SARS-CoV-2 can be transmitted effectively among humans and exhibits high potential for a pandemic [5, 10, 13]. It is very important to stay informed during this outbreak. Moreover, this novel virus is new to the scientific world and many features of the virus are still not understandable due to its new strains [14]. Hence, the worldwide researchers are now very active to explore the new insights of the virus in order to understand its biological character and mode of spreading. This real boost of research interest on the virus has actually started after the emergence of SARS and MARS, and subsequent COVID-19.

II. Materials and Methods

In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 200,000 scholarly articles, including over 90,000 with full text, about COVID-19, SARS-CoV-2, and related corona viruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. [15].

III. Exploratory Data Analysis (EDA)

The dataset covers 163 countries and almost 2 full months from 2020, which is enough data to get some clues about the pandemic. Let's see a few plots of the worldwide tendency in Fig. 1 to extract some insights:

Observations:

The global curve shows a rich fine structure, but these numbers are strongly affected by the vector zero country, China. Given that COVID-19 started there, during the initial expansion of the virus there was no reliable information about the real infected cases. In fact, the criteria to consider infection cases was modified around 2020-02-11, which strongly perturbed the curve as you can see from Fig. 1.

A. COVID-19 Behavior

Since China was the initial infected country, the COVID-19 behavior is different from the rest of the world. The medical system was not prepared for the pandemic; in fact no one was aware of the virus until several cases were reported.

Moreover, China government took strong contention measures in a considerable short period of time and, while the virus is widely spread, they have been able to control the increasing of the infections.

Observations:

a) Smoothness: Both plots are less smooth than theoretical simulations or the curve from the rest of the world cumulative.

b) Infected criteria: The moment in which the criteria to consider an infected case was changed is directly spotted.

c) Irregularities: There are some irregularities. I should check the literature in depth to look for evidences, but the reasons may be that both the resources spent to monitor the epidemic and the security measures that have been changing over time.

d) Plateaux: It looks like the curve has reached a plateaux, which would imply that China is on their maximum of contagion, which strongly perturbed the curve as you can see from Fig. 2.

B. Italy, Spain, UK and Singapore

Both Italy and Spain are experiencing the larger increase in COVID-19 positives in Europe. At the same time, UK is a unique case given that it's one of the most important countries in Europe but recently has left the European Union, which has create an effective barrier to human mobility from other countries. The fourth country studied in this section is Singapore, since it's an Asiatic island, is closer to China and its socio-economic conditions is different from the other three countries, which strongly perturbed the curve as you can see from Fig. 3.

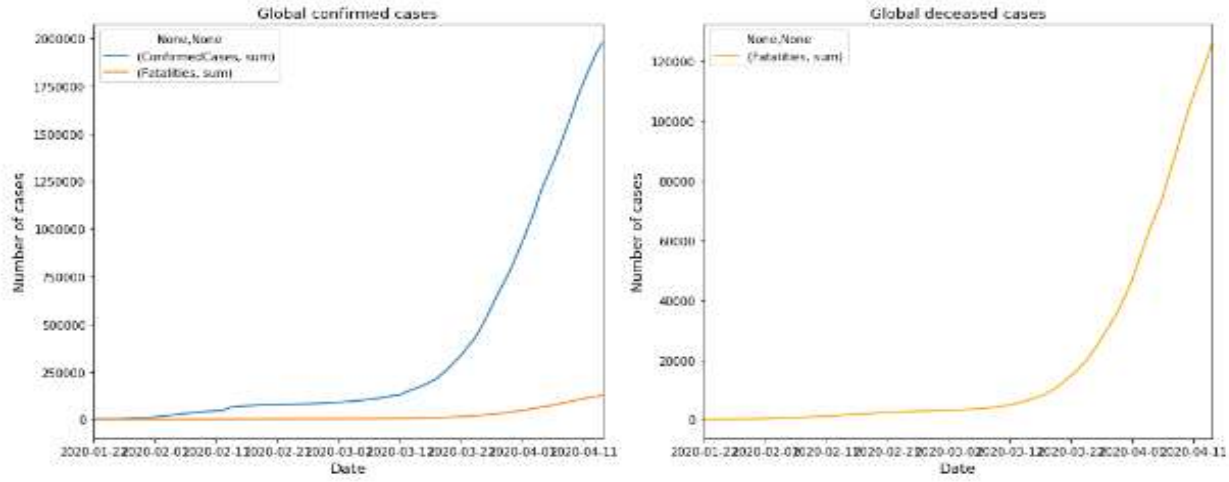

Fig. 1. Global Confirmed Cases Date-Wise.

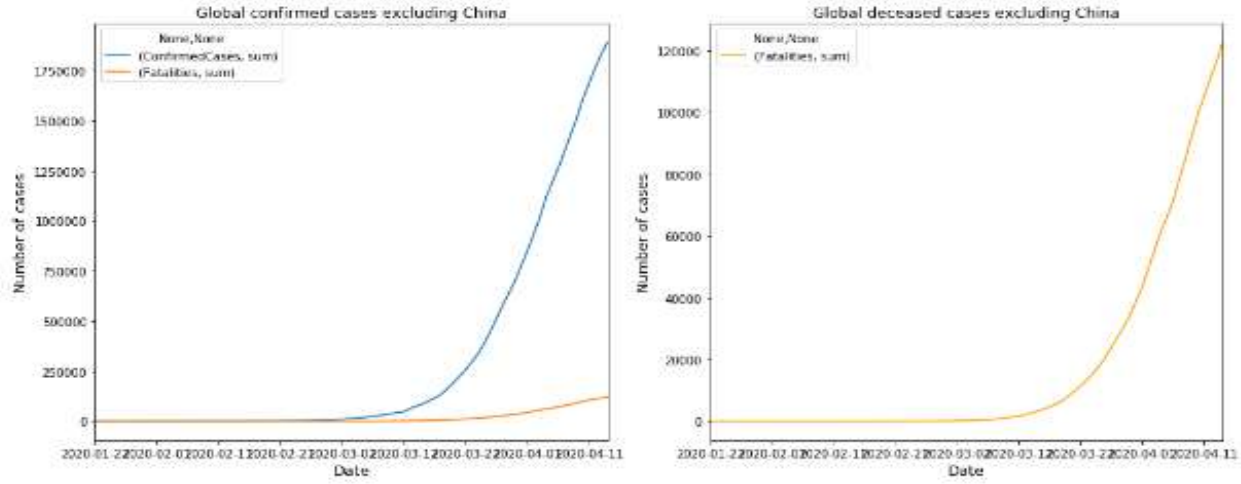

Fig. 2. Global Confirmed Cases Excluding China Date-Wise.

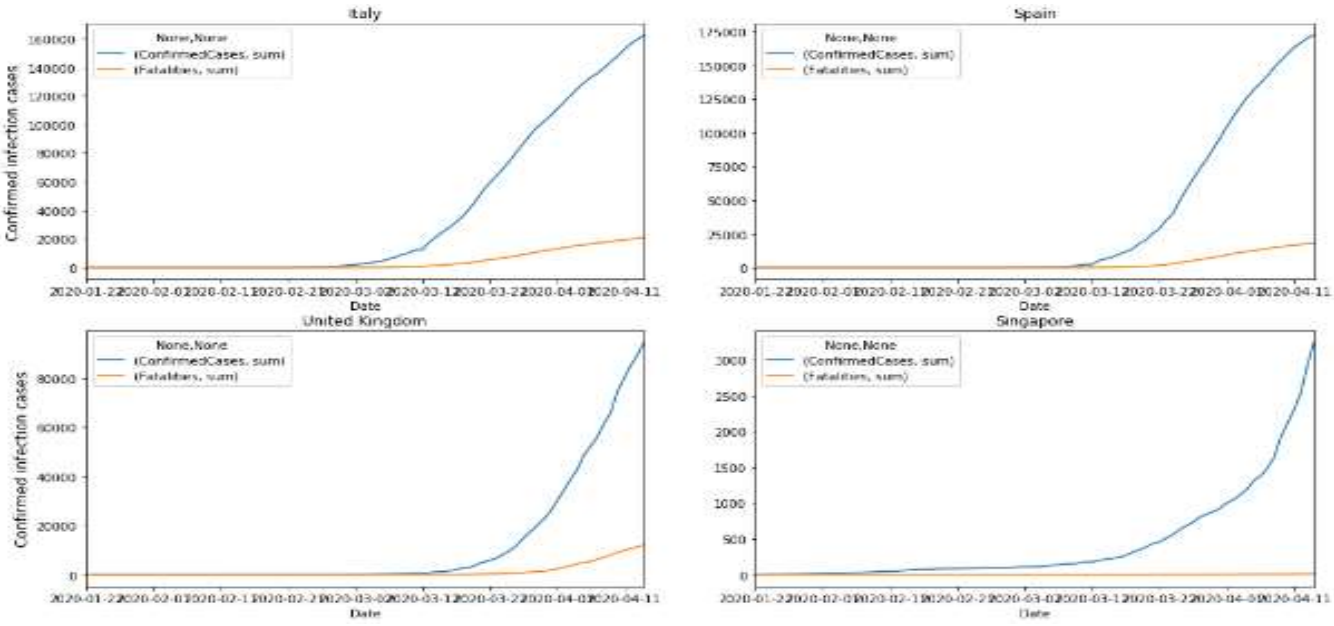

Fig. 3. Confirmed Infection Cases in different Countries.

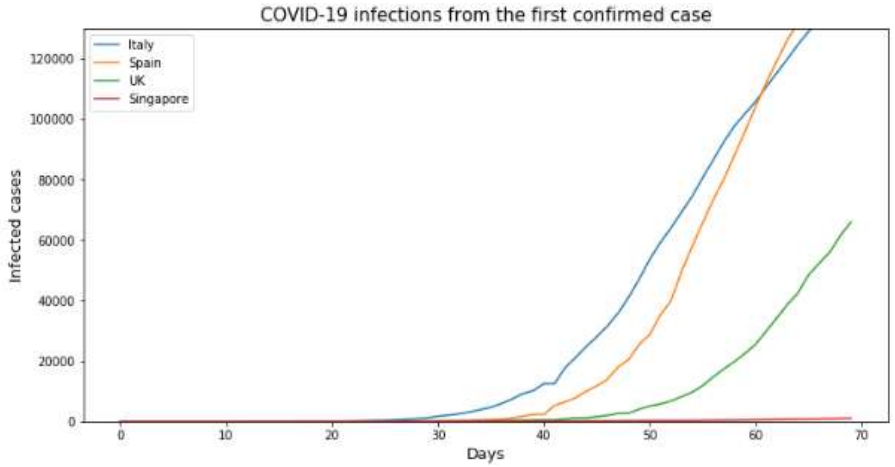

Fig. 4. Infections from the First Confirmed Case Spread Over different Countries.

As a fraction of the total population of each country, in order to compare the four countries, it's also interesting to see the evolution of the infections from the first confirmed case, which is plotted in Fig. 4.

Observations:

a) Italy. With almost 120.000 confirmed cases, Italy shows one of the most alarming scenarios of COVID-19. The infections curve is very steep, and more than 2% of population has been infected.

b) Spain. Spain has the same number of cumulative infected cases than Italy, near 120.000. However, Spain's total population is lower (around 42 millions) and hence the percentage of population that has been infected rises up to 3%.

c) United Kingdom. Despite not being very far from them, the UK shows less cases. This may be due to the number of tests performed, but it's soon to know for sure. The number of cases is around 40.000, this is, a 0.6 of the total population.

d) Singapore. Singapore is relatively isolated given that is an island, and the number of international travels is lower than for the other 3 countries. The number of cases is still very low (>1000), despite the general tendency is to increase. However, the infections started faster in the beginning, but the slope of the infections curve hasn't increased very much in the past weeks. A 0.2% of the population was infected.

IV. SIR Model

Some general behavior of the virus in aggregated data, for the country where the corona virus was originated and for four other interesting countries. The purpose of this study is to develop a predictive model in order to understand the key factors that impact the COVID-19 transmission, Let's move on to one of the most famous epidemiologic models: SIR, the workflow shown in Fig. 5.

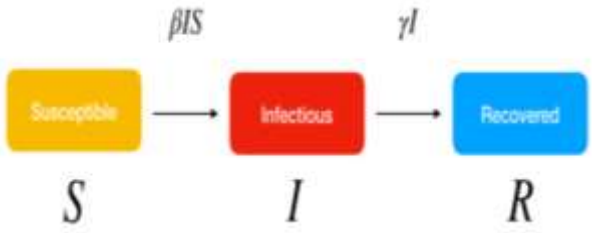

Fig. 5. SIR Workflow.

SIR is a simple model that considers a population that belongs to one of the following states:

- *Susceptible (S).* The individual hasn't contracted the disease, but she can be infected due to transmission from infected people.
- *Infected (I).* This person has contracted the disease.
- *Recovered/Deceased (R).* The disease may lead to one of two destinies: either the person survives, hence developing immunity to the disease, or the person is deceased.

There are many versions of this model, considering birth and death (SIRD with demography), with intermediate states, etc. However, since world is in the early stages of the COVID-19 expansion and interest is focused in the short term, will consider that people develops immunity (in the long term, immunity may be lost and the COVID-19 may come back within a certain seasonality like the common flu) and there is no transition from recovered to the remaining two states.

A. Implementing the SIR Model

SIR model can be implemented in many ways: from the differential equations governing the system, within a mean field approximation or running the dynamics in a social network (graph). For the sake of simplicity run a numerical method (Runge-Kutta) to solve the differential equations system.

In order to solve the differential equations system, a 4rth order Runge-Kutta method is developed.

And finally, to obtain the evolution of the disease, simply define the initial conditions and call the Runge-Kutta method.

The number of infected cases increases for a certain time period, and then eventually decreases given that individuals recover/decease from the disease. The susceptible fraction of population decreases as the virus is transmitted, to eventually drop to the absorbent state 0, which is predicted in Fig. 6. The opposite happens for the recovered/deceased case. Notice that different initial conditions and parameter values will lead to other scenarios, feel free to play with these numbers to study the system.

B. Fit SIR Parameters to Real Data

The SIR model is purely theoretical, and interested in a real approximation of the COVID-19 expansion in order to extract insights and understand the transmission of the virus. Model needs to extract the ββ and γγ parameters for each case to predict the evolution of the system.

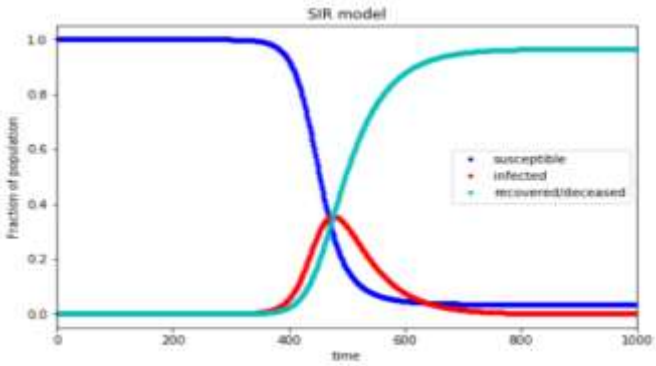

Fig. 6. Susceptible Fraction of Population Decreases as the Virus is Transmitted.

C. Data Enrichment

Analyzing SIR simulations was meant to understand a model that approximately resembles the transmission mechanism of many viruses, including the COVID-19. However, there are alternative methods that may prove being equally useful both to predict and to understand the pandemic evolution. Many of these methods rely on having rich data to extract conclusions and allow algorithms to extrapolate patterns in data, and that is exactly what is going to be implemented.

D. Main Workflow of this Section

- Join data, filter dates and clean missing.
- Compute lags and trends.
- Add country details.

Disclaimer: This data enrichment is not mandatory and could end up without using all of the new features in the model. However, this is consider as a didactical step that will surely add some value, for example in an in-depth exploratory analysis.

1) Join data, filter dates and clean missing: First of all, let's perform some pre-processing to prepare the dataset, consisting on:

- *Join data.* Join train/test to facilitate data transformations.
- *Filter dates.* According to the challenge conditions, remove Confirmed Cases and Fatalities post 2020-03-12. Create additional date columns.
- *Missing.* Analyze and fix missing values.

Observations:

a) "Confirmed Cases" and "Fatalities" are now only informed for dates previous to 2020-03-12.

b) The dataset includes all countries and dates, which is required for the lag/trend step.

c) Missing values for "Confirmed Cases" and "Fatalities" have been replaced by 0, which may be dangerous if it is not remembered at the end of the process. However,

since training is done only on dates previous to 2020-03-12, this won't impact the prediction algorithm.

d) A new column "Day" has been created, as a day counter starting from the first date.

2) Compute lags and trends: Enriching a dataset is a key to obtain good results. In this case, two different transformations are applied:

a) Lag. Lags are a way to compute the previous value of a column, so that the lag 1 for Confirmed Cases would inform the column from the previous day.

b) Trend. Transforming a column into its trend gives the natural tendency of this column, which is different from the raw value.

The backlog of lags is applied for 14 days, while for trends is for seven days.

3) Add country details: Variables like the total population of a country, the average age of citizens or the fraction of people living in cities may strongly impact on the COVID-19 transmission behavior. Hence, it's important to consider these factors. The dataset is based on Web Scrapping for this purpose.

4) Predictions for the early stages of the transmission: The objective in this section consists of predicting the evolution of the expansion from a data-centric perspective, like any other regression problem. To do so, remember that the challenge specifies that submissions on the public LB should only contain data previous to 2020-03-26.

a) Tools utilized: Previously published automated machine learning tool (https://automated machinelearning-gitamcse.shinyapps.io/MLPv3/) [16, 17] is utilized here for building multiple models on the imputed dataset. The natural advantage of the AMLT tool is to choose multiple train and test sets coupled with a suitable statistical algorithm to build the best models out of the available data. AMLT tool also does the test validation automatically, which will be helpful to understand the accuracy of each model.

b) Models to apply:

1) Linear Regression for one country
2) Linear Regression for all countries

V. Linear Regression for One Country

Since we are interested into predicting the future time evolution of the pandemic, the first approach consists on a simple Linear Regression. However, remind that the evolution is not linear but exponential (only in the beginning of the infection), so that a preliminary log transformation is needed.

Visual comparison of both cases for Spain and with data from last 10 days informed, starting on March 1st is depicted in Fig. 7.

As you see, the log transformation results in a fancy straight-like line, which is awesome for Linear Regression. However, let me clarify two important points:

- This "roughly exponential behavior" is only true for the initial infection stages of the pandemic (the initial increasing of infections on the SIR model), but that's exactly the point where most countries are at the moment.
- Why do I only extract the last 10 days of data? For three reasons:

1) In order to capture exactly the very short term component of the evolution

2) To prevent the effects of certain variables that have been impacting the transmission speed (quarantine vs. free circulation)

3) To prevent differences on criteria when confirming cases (remember that weird slope on the China plot?).

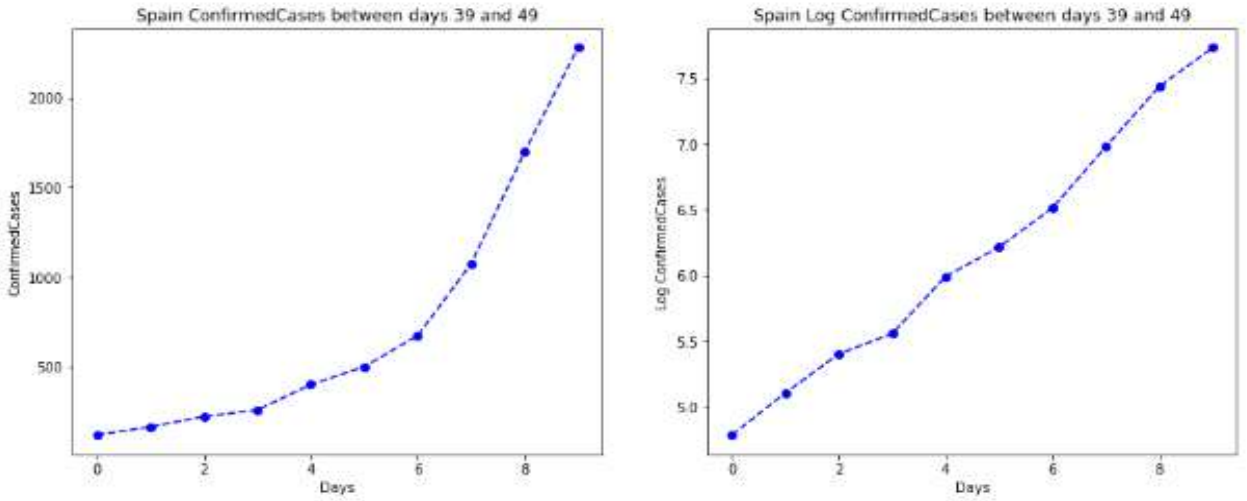

Fig. 7. Spain Confirmed Cases Day-Wise.

This first model is very simple, and only elemental features will be considered: Country/Region, date information, Long and Lat. Lags. Engineered columns like lags, trends and country details are not introduced as an input. Finally, the workflow for the Basic Linear Regression model is:

1) Features. Select features.

2) Dates. Filter train data from 2020-03-01 to 2020-03-18.

3) Log transformation. Apply log transformation to Confirmed Cases and Fatalities.

4) Infinites. Replace infinites from the logarithm with 0. Given the asymptotic behavior 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, but still needs to be tackled sooner or later in order to obtain a clean solution.

5) Train/test split. Split into train/valid/test.

6) Prediction. Linear Regression, training country by country and joining data.

7) Submit. Submit results in the correct format, and applying exponential to reverse log transformation.

A. Linear Regression for All Countries

An alternative method to setting the number of days for the training step is to simply keep all data for each country since the first case was confirmed. However, since there are certain countries were the initial outbreak was very smooth (i.e. in Spain there was only one confirmed case for 7 days in a row), predictions maybe biased by these initial periods.

Final LMSE score for week 2, with training data prior to 2020-03-19 and measures on date 2020-04-01: 1.19681.

VI. Conclusion

A. Results

1) Parameters. Two full weeks of training used (from February 26th to March 11th), with their previous 30 lags.

2) Enough data. (Spain, Italy, Germany). For countries with several Confirmed Cases!=0 in the train dataset (prior to March 11th), predictions are very precise and similar to actual confirmed data.

3) Poor data. Countries with a small number of data points 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.

4) No data. When the number of confirmed cases in the train dataset is 0 or negligible, the model predicts always no infections.

B. Discussion

1) The objective of this work is to provide some insights about the COVID-19 transmission from a data-centric perspective in a didactical and simple way. Predicted results should not be considered in any way an affirmation of what will happen in the future. Observations obtained from data exploration are personal opinions.

2) Models tailored specifically for epidemic spreading (i.e. SIR and its versions) are designed to reproduce a certain phenomenology, in order to understand the underlying mechanics of a contagion process. On the other hand, the simple machine learning approaches I used aim to predict the short term evolution of the infection in the current regime. They might eventually help to find some features or parameters that are particularly important for the model's fitting, but by no means should they be confused with scientific epidemic models.

3) The success of the current predictions is strongly dependent on the current spreading regime, in which the number of infections is still increasing exponentially for many countries. However, they cannot provide a reliable expected day by which the maximum contagion peak will be reached. Epidemic models are closer to obtaining such estimations, but there's a large number of variables that need to be considered for this (quarantines, quality of the medical resources deployed, environmental measures…).

4) In order to achieve such results, a considerable amount of tuning is required. Filter how many previous dates should be used for the fitting step, when to use lags or not, and even missing replacements were very rough due to the log transformation.

C. Declaration

Predictive models can be used for several purposes, but they never (try to) substitute recommendations from experts.

References

[1] Lu H, Stratton CW, Tang YW. Outbreak of pneumonia of unknown etiology in Wuhan China: the mystery and the miracle. J Med Virol 2020 Jan 16 [Epub ahead of print]. doi: 10.1002/jmv.25678.

[2] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 2020 Jan 29 [Epub ahead of print]. doi: 10.1056/NEJMoa2001316.

[3] Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. Severe acute respiratory syndrome-related coronavirus: the species and its viruses—a statement of the Coronavirus Study Group. bioRxiv 2020 Feb 11. doi: 10.1101/2020.02.07.937862.

[4] Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020;395:507–13. doi: 10.1016/ S0140- 6736(20)30211- 7.

[5] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395:497–506. doi: 10.1016/S0140-6736(20)30183-5.

[6] Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet 2020;395:470–3. doi: 10.1016/S0140-6736(20)30185-9.

[7] Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First case of 2019 novel coronavirus in the United States. N Engl J Med 2020 Jan 31 [Epub ahead of print]. doi: 10.1056/NEJMoa2001191.

[8] Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA 2020 Feb 7 [Epub ahead of print]. doi: 10.1001/jama.2020. 1585.

[9] Chang D, Lin M, Wei L, Xie L, Zhu G, Dela Cruz CS, et al. Epidemiologic and clinical characteristics of novel coronavirus infections involving 13 patients outside Wuhan, China. JAMA 2020 Feb 7 [Epub ahead of print]. doi: 10.1001/ jama.2020.1623.

[10] Carlos WG, Dela Cruz CS, Cao B, Pasnick S, Jamil S. Novel Wuhan (2019- nCoV) coronavirus. Am J Respir Crit Care Med 2020;201:P7–8. doi: 10.1164/ rccm.2014P7.

[11] Zhao S, Lin Q, Ran J, Musa SS, Yang G, Wang W, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak. Int J Infect Dis 2020;92:214–17. doi: 10.1016/j.ijid.2020.01.050.

[12] Biscayart C, Angeleri P, Lloveras S, Chaves T, Schlagenhauf P, Rodriguez- Morales AJ. The next big threat to global health? 2019 novel coronavirus (2019-nCoV): What advice can we give to travellers? — Interim recommen- dations January 2020, from the Latin-American Society for Travel Medicine (SLAMVI). Travel Med Infect Dis 2020:101567. doi: 10.1016/j.tmaid.2020.101567.

[13] Munster VJ, Koopmans M, van Doremalen N, van Riel D, de Wit E. A novel coronavirus emerging in China—key questions for impact assessment. N Engl J Med 2020 Jan 24 [Epub ahead of print]. doi: 10.1056/NEJMp20 0 0929.

[14] Qing, E., Gallagher, T., 2020. SARS coronavirus redux. Trends Immunol. 41, 271–273. https://doi.org/10.1016/j.it.2020.02.007.

[15] https://www.semanticscholar.org/cord19: arXiv:2004.10706.

[16] DeepaRani Gopagoni, P V Lakshmi, 2020. Automated machine learning tool, the first stop for data science and statistical model building. IJACSA 2020, 11(2),410-418,DOI: 10.14569/IJACSA.2020.0110253. https://automatedmachinelearning-gitamcse.shinyapps.io/MLPv3/.

[17] DeepaRani Gopagoni, P V Lakshmi, An Application of Machine Learning Strategies to Predict Alzheimer's Illness Progression in Patients International Journal of Advanced Research in Engineering and Technology (IJARET) Volume 11, Issue 6, June 2020, pp. 1056-1063, Article ID: IJARET_11_06_095 DOI: 10.34218/IJARET.11.6.2020.95.