MLE for Inverse problem of SEIR for COVID-19 for world

Abstract

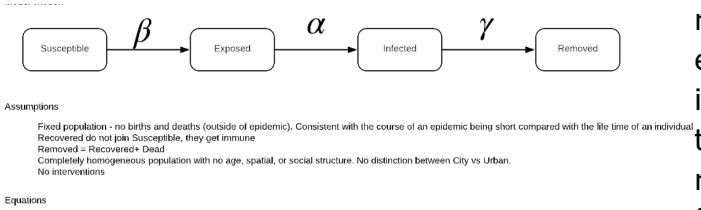
The SEIR model is used to study the spread of epidemics. The parameters here are assumed to be specific to not only the disease but the country/region. We have applied ML techniques to solve the inverse problem for epidemics based on data, deviating from traditional approaches and stochastic models. Since real situations have far more variables, our new approach is amenable to extension by adding more aspects as needed. Our algorithm uses a multi-step SEIR model approach with the state-space search for parameters.
The data is taken from publicly available data sources (primarily Kaggle). We present a Machine Learning based algorithm to estimate parameters with windowing. A custom accuracy score is used to choose the right window for estimating parameters. Time/Window dependent parameter (alpha, beta, and gamma) values present understanding of measures taken by countries and is a learning that can be shared in real-time. All the code in the literate programming paradigm is in the public domain.

Introduction

COVID-19 happened too fast for countries to react. Many lives could have been saved if the response was quicker and we had the ability to learn from each other. Also, we had no way of measuring how a country was doing and whether those measures were working. The best view we have been getting is through graphs, where we look for peaks visually. Beyond the fact that we have peaked or not, we cannot derive much information from these graphs. To have a real-time measurement and sharing of learning is the goal with which we started an effort to fit the SEIR model to data. By reducing the data to 3 parameters (alpha, beta, gamma), possibly as a time series, we can quantify the impact of real-world events like social distancing, quarantine, social events, etc.. With these goals, we started a public code repository of ML Jupyter notebooks - https://github.com/baladutt/modeling-epidemics-COVID-19. All the code mentioned in this paper is present and is in the public domain.

SEIR model

SEIR stands for Susceptible, Exposed, Infected, Recovered. The compartmental model classifies the population into four mutually exclusive groups: susceptible (at risk of contracting the disease), exposed (infected but not yet infectious), infectious (capable of transmitting the disease), and removed (those who recover or die from the disease).



$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

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

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \alpha E$$

$$\frac{dI}{dt} = \alpha E - \gamma I$$

$$N = S + E + I + R$$

$$\beta = contact\ rate$$

$$\alpha = \frac{1}{T_{incubation}}$$

$$\gamma = \frac{1}{T_{infection}}$$

$$R_0 = \frac{\beta}{\gamma}$$

The ODE equations are parameterized by alpha, beta, and gamma and provide temporal evolution. Simulation is done using Euler's method of differential equations.

The simulation works with the time-series values of Gamma, assuming a changing Gamma due to say better treatment, etc.

The simulation starts with 1 person being infected. More people entering the country infected may be captured by changing alpha and beta.

Related Work

A simple concise introduction to SIR model is presented here with more references to go into details.[1] A simulation and deeper introduction to the model are provided by Harri in his popular article[2]. Henri extends the SIR model to include deaths, recovered and uses transition probabilities to simulate[3]. The article goes on to include resources and age-dependent fatalities. The code is available with an approach to extend the general framework to add more compartments. Williams presents an approach to get the COVID vulnerability index to predict most

[1] https://mathworld.wolfram.com/Kermack-McKendrickModel.html
[2] https://www.washingtonpost.com/graphics/2020/world/corona-simulator/
[3] Infectious Disease Modelling: Beyond the Basic SIR Model - https://towardsdatascience.com/infectious-disease-modelling-beyond-the-basic-sir-model-216369c584c4

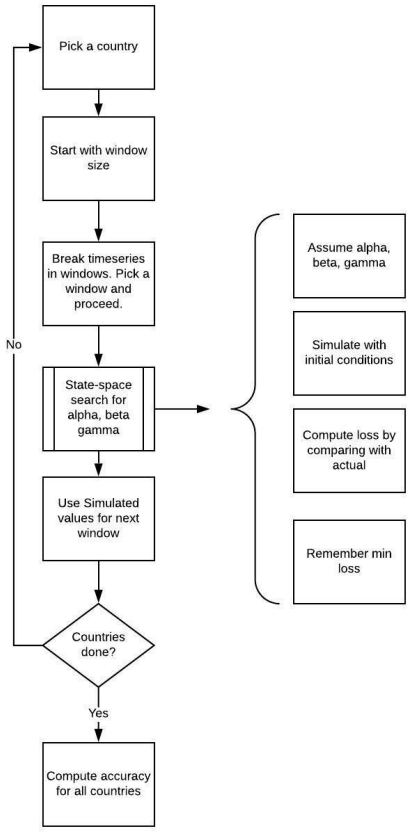
vulnerable before testing.[4] Meinhard compares countries on how similar they were in response to COVID-19[5]. The article continues to cluster countries, however, the approach is custom. Mary presents the use of an R0(t), the value of R0 changing with time t[6]. Mohak presents options for a country with simulation and learning from other countries.[7] The inverse problem of finding parameters is a hard problem with no unique solution.[3]

Our work

Data

The data is taken from publicly available data sources (primarily Kaggle).

World	https://github.com/CSSEGISandData/COVID-19.git
US	https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset
India	https://www.kaggle.com/sudalairajkumar/covid19-in-india
Italy	https://www.kaggle.com/sudalairajkumar/covid19-in-italy
Population	https://www.kaggle.com/tanuprabhu/population-by-country-2020



Algorithm

The core of the algorithm is a Simulation model. This simulation model is based on the SIER equations described earlier. It takes a set of 3 parameters (namely, alpha, beta, gamma)and an initial set of SEIR values. It computes the SEIR values over a time period.
After the simulation, loss is computed as a weighted sum of squared errors of simulated vs actual of infected and removed. The goal of the inverse problem is to minimize the loss. Errors of removed are given a higher weight as the numbers are much smaller than infected.
This simulation and loss is used to find a solution for inverse problem. So, an optimum value of

rofiles/blogs/building-a-covid-19-vulnerability-index
19 -
ntries-react-similar-to-covid-19-machine-learning-provides-the-answer-

.com/1834700/rt-the-real-time-r0-guiding-how-to-lift-coronavirus-lockd
D&_js_v=0.1#referrer=https%3A%2F%2Fwww.google.com&_tf
3A%2F%2Fqz.com%2F1834700%2Frt-the-real-time-r0-guiding-how-to-

[7] https://medium.com/@mohakgupta_55841/coronavirus-in-india-make-or-break-5a13dfb9646d

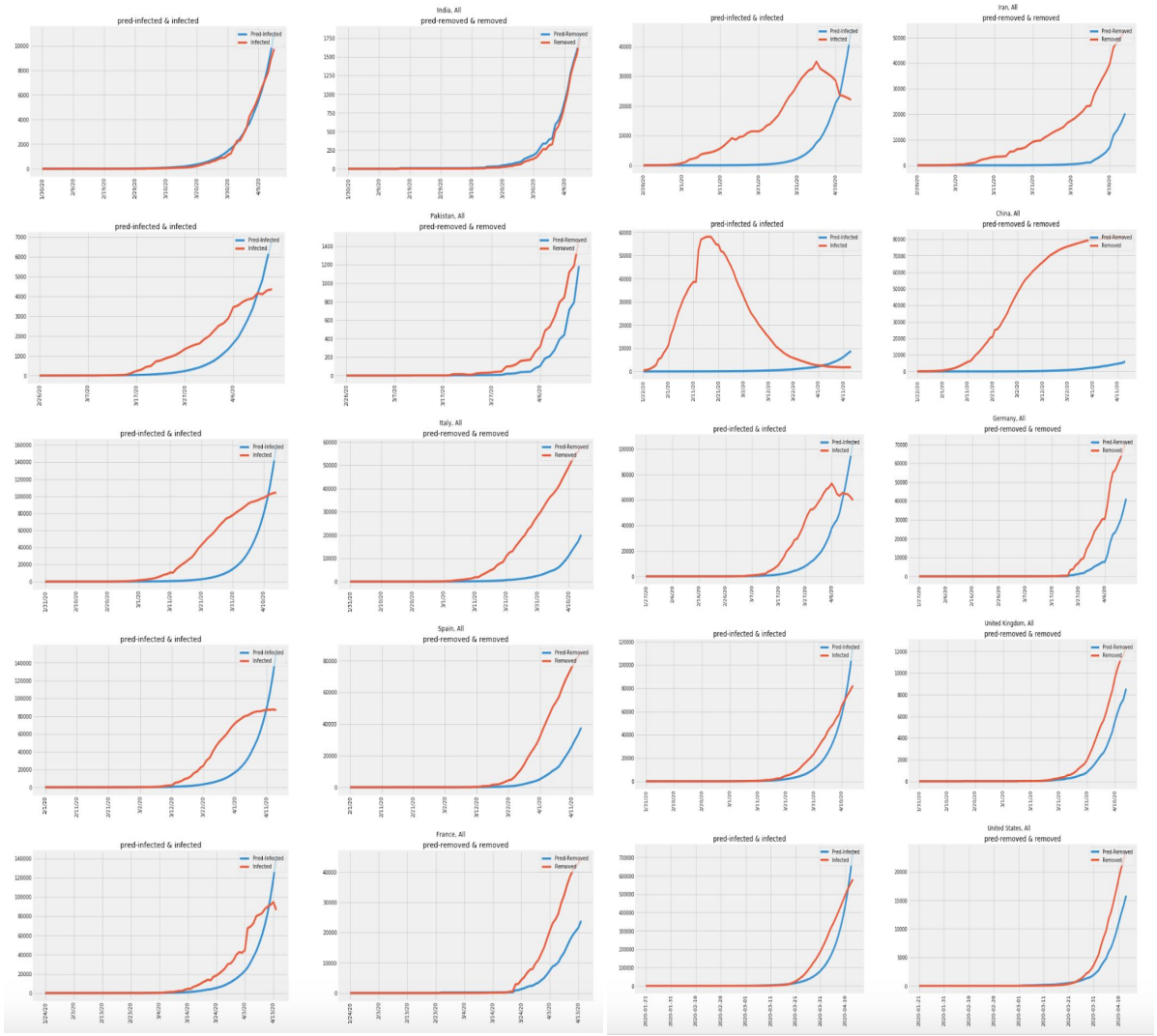
parameters is obtained by doing a state space search. Gammas were tried with multiple options like searching for them, using a mean computed value given from actual data or using a time series value from actual data. This way the algorithm fits a given time series of infected and removed to alpha, beta and gamma values.

Now the data for a country over total time is broken into windows. For each window the values of these parameters are computed incrementally which helps in the prediction of subsequent values. It is essentially a feed forward mechanism of prediction. This results in alpha, beta and gamma as time-dependent parameters.

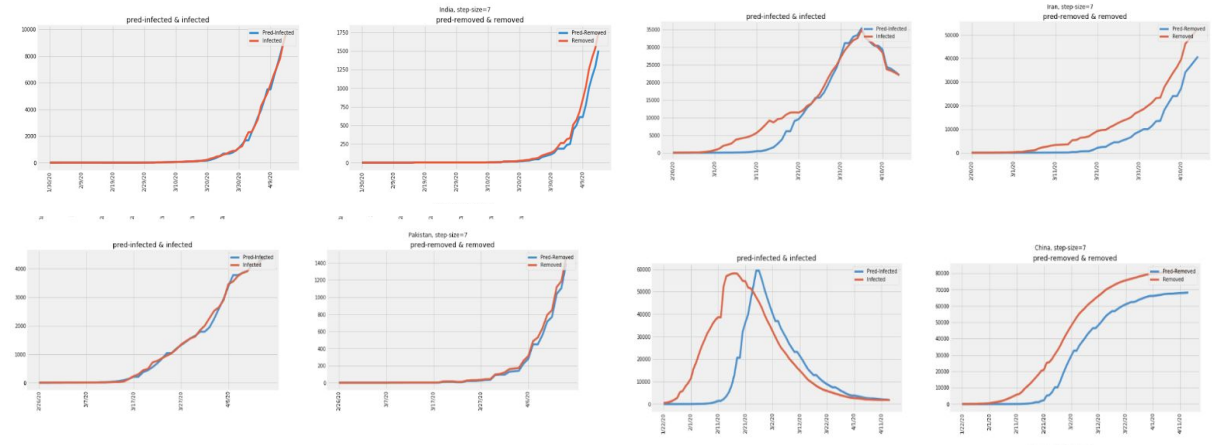
Time series for each country is tried over multiple window sizes to find a window that gives the best results. Finally, a custom accuracy score was defined to select the best window. The accuracy score uses worst-case country error for a time window as a representative of the window's performance. And the window which gives the least worst-case performance is chosen.

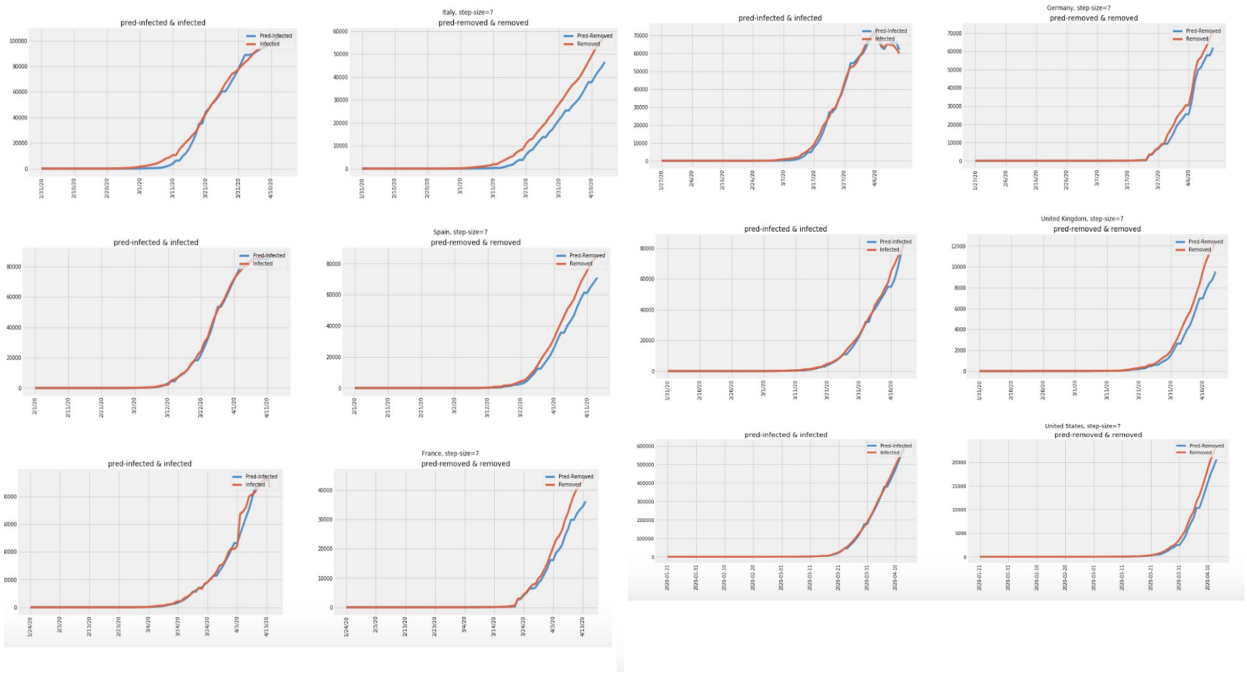
Results and observations

Window Size = All

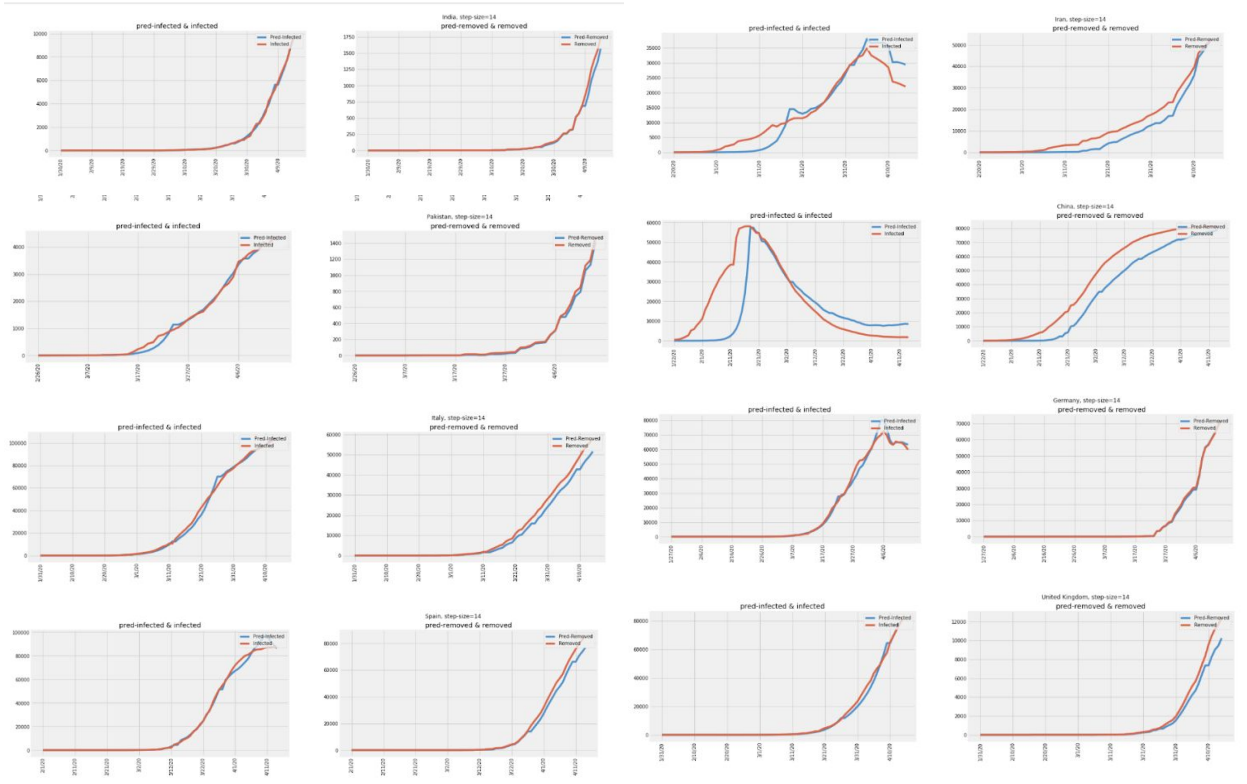


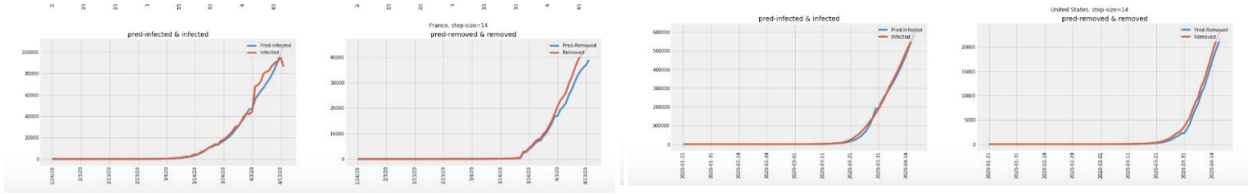
Window Size = 7



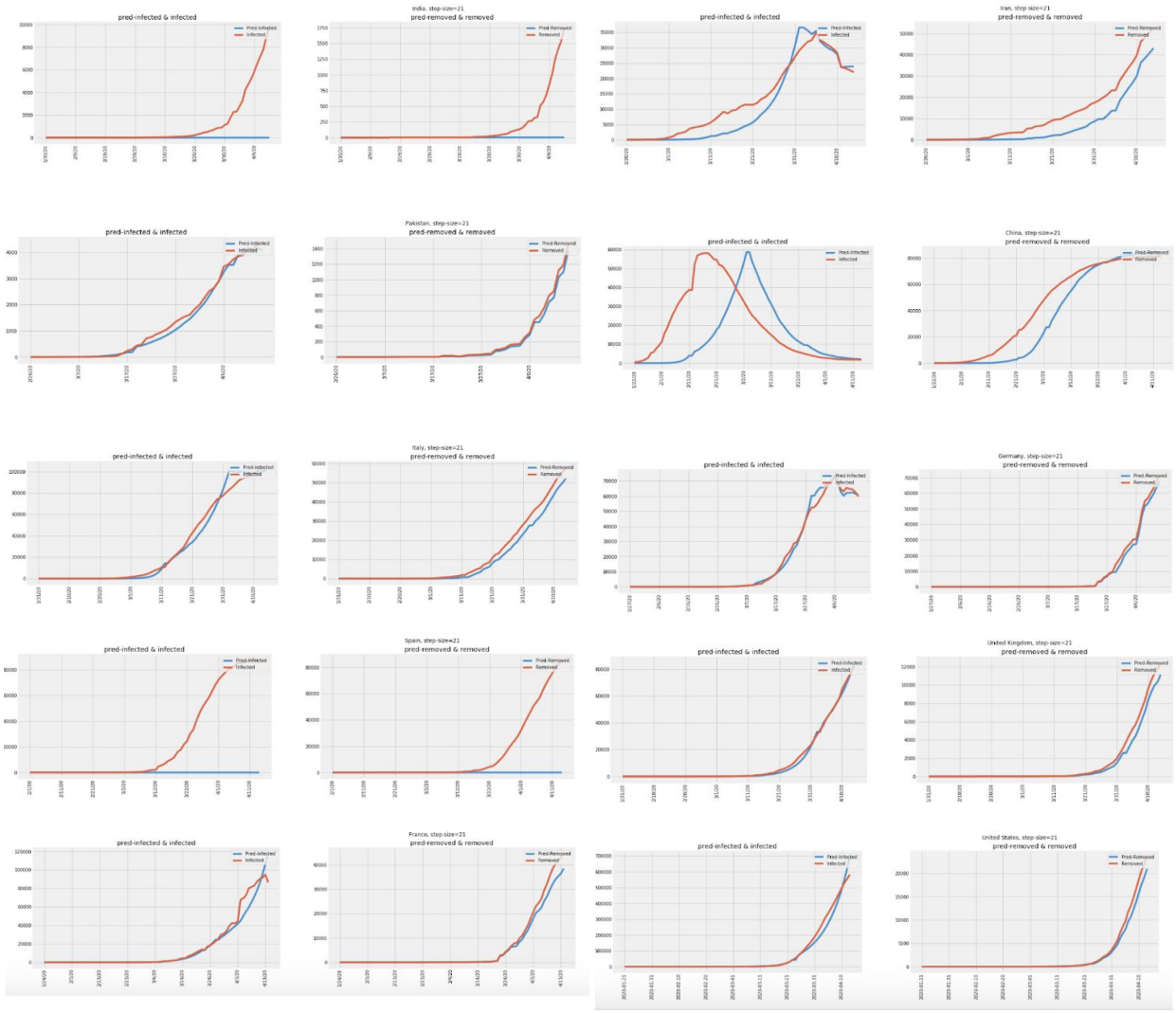


Window size = 14





Window size = 21

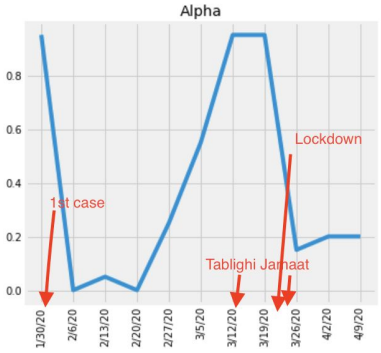


Measure of accuracy

Window Size	Accuracy score	Explanation: We find that window 14 did the best. While small windows are intuitively expected to fit the data better,
All	0.100000	

7	0.455602	they may cause overfitting. Also, the inherent latencies in epidemics favor optimum window sizes.
14	0.555973	
21	0.163536	

Conclusion and Future work



How countries have fared can be seen by the change in alpha, beta and gamma parameters. These changes can be correlated to actions and events in the real world. An example of such mapping is presented here for country India.
These can be learnings for all. Also, one can see the quantitative effects of these actions/events. We expect that in future epidemics such real-time learning can save much of the loss of human life, pain and material impact.

Future work would include deaths and other states in the model. Also, stochastic modeling could be compared. Countries could be clustered based on parameters and the cluster may be studied. It is possible that we learn more about the disease from clusters, for example, how does the temperature, humidity, culture, demographics of the countries affect.

References

[1] Kermack W, McKendrick A (1927) Contributions to the mathematical theory of epidemics—I. Proc R Soc Edinb A 115:700–721

[2] arXiv:2004.07738
[q-bio.PE]
@misc{giudici2020inversion,
title={Inversion of a SIR-based model: a critical analysis about the application to COVID-19 epidemic},
author={Mauro Giudici and Alessandro Comunian and Romina Gaburro},
year={2020},
eprint={2004.07738},
archivePrefix={arXiv},
primaryClass={q-bio.PE}
}

[3] Tchavdar T. Marinov, Rossitza S. Marinova, Joe Omojola, Michael Jackson,

Inverse problem for coefficient identification in SIR epidemic models,
Computers & Mathematics with Applications,
Volume 67, Issue 12,
2014,
Pages 2218-2227,
ISSN 0898-1221,
https://doi.org/10.1016/j.camwa.2014.02.002.
(http://www.sciencedirect.com/science/article/pii/S0898122114000601)

[4] Hyuk-Jun Chang. (2017). Estimation of basic reproduction number of the middle east respiratory syndrome coronavirus (MERS-CoV) during the outbreak in south korea, 2015. *Biomedical Engineering Online, 16* doi:http://dx.doi.org/10.1186/s12938-017-0370-7

https://www.indiatoday.in/india/story/timeline-of-nizamuddin-markaz-event-of-tablighi-jamaat-in-delhi-1661726-2020-03-31
https://trends.google.com/trends/?geo=US
https://en.wikipedia.org/wiki/2020_coronavirus_lockdown_in_India