Global Spread of COVID-19 Prediction

Kanak Choudhury

PhD3

kanakc@iastate.edu

Department of Statistics

Statistical Machine Learning, SVR

Mingdian Liu

PhD3

mingdian@iastate.edu

Department of Electrical and Computer Engineering

Photonics, Biosensor

Submitted To

Prof. Christopher Quinn

Assistant Professor

Course Instructor: ComS 574

Department of Computer Science

1 Introduction

The whole world is suffering for the Covid-19 pandemic. It becomes the major concern for every country to predict number of deaths and infected persons under different conditions and measures. In this project, prediction of the number of daily deaths has been considered using some machine learning models. It is important to mention that for any machine learning model, we need lots of historic information that can be used to fit a model and using that model we can predict for the future. However, in this type of prediction we do not have any previous information about number of daily deaths or infected cases and related measures applied to mitigate or reduce number of deaths. What we have, are all continuously applied measures and information that makes the difficulty to apply machine learning models directly. There are many other methods (e.g. different pandemic prediction models like SIR, SIRD) that can be used to predict deaths and infected numbers though there are some limitations for such models. For this reason, I have considered two step prediction model that can incorporate different measures implemented to control this covid-19 pandemic as well as country specific economic, health and demographic indicators. In the first set, I have used popular methods (SIR, SIRD) generally used to predict deaths and infected numbers and in the second step, using these prediction as well as other country specific indicators and some factors generated based on some conditions were used to fit machine learning models to predict number of daily cumulative deaths.

2 Related Work

There are different types of epidemiological models used to predict susceptible, previously unexposed by the disease, infected, currently exposed by the disease, and recovered, successfully recovered by the disease, and deaths, died by the disease. The most common and simple model that has been used is called S-I-R model proposed by (Kermack & McKendrick, 1927) and this model does not depends on time or other factors such as any measures taken to control disease. Different modified models including time dependent models have been proposed by authors (Keeling & Rohani, 2008), (Amaro, 2020), (Calafiore, Novara, & Possieri, 2020), (Chen, Lu, & Liu, 2020), (Bayes, Rosas, & Valdivieso, 2020), (Harko, Lobo, & Mak, 2014), (Johnson). Applications of these models on Covid-19 pandemic have been studied by authors. (Website 2. , 2019) (Dhanwant & Ramanathan, 2020). (Fanelli & Piazza, 2020) analyzed the temporal dynamics of the coronavirus disease 2019 outbreak in China, Italy, and France in the time window 22/01 − 15/03/2020. We study the increases of infections and deaths in Sweden caused by Covid-19 with several different models: (Qi, Karlsson, Sallmen, & Wyss, 2020) studied on Covid-19 Using susceptible-infected (SI) model and the standard SIR model. They also studied susceptible-infected-deceased (SID) correlations based on SIR model.

The Institute for Health Metrics and Evaluation (IHME) developed mixed effects non-linear regression framework to estimate the trajectory of the cumulative and daily death rate as a function of the implementation of social distancing measures, supported by additional evidence from mobile phone data (IHME Team, 2020).

3 Background

3.1 Susceptible-Infected-Recovered (SIR) Model

McKendrick and Kermack proposed a model called the Susceptible-Infected-Recovered (SIR) model to describe the spread of diseases (Kermack & McKendrick, 1927). It was represented by the following nonlinear system of ordinary differential equations-

$$\frac{dx}{dt} = -\beta x(t)y(t) \tag{1}$$

$$\frac{dy}{dt} = \beta x(t)y(t) - \gamma y(t) \tag{2}$$

$$\frac{dz}{dt} = \gamma y(t) \tag{3}$$

where $x(t)$ is the susceptible (S), $y(t)$ is the infected (I) and $z(t)$ is the recovered (R) and at time 0, $x(0) = N_1 \geq 0$, $y(0) = N_2 \geq 0$ and $z(0) = N_3 \geq 0$, $\beta \geq 0$ is the infection rate and $\gamma \geq 0$ is the mean recovery rate for a fixed population, N such that $N_1 + N_2 + N_3 = N$ (Harko, Lobo, & Mak, 2014).

3.2 Susceptible-Infected-Recovered-Death (SIRD) Model

In SIRD model, one extra differential equation has been considered to find the rate of death and are given by-

$$\frac{dS}{dt} = -N^{-1}\beta SI \tag{4}$$

$$\frac{dI}{dt} = N^{-1}\beta SI - (\gamma + \alpha)I \tag{5}$$

$$\frac{dR}{dt} = \gamma I \tag{6}$$

$$\frac{dD}{dt} = \alpha I \tag{7}$$

where S, I, R, and D represent susceptible, infected, recovered, fatal, respectively. In this model, α represents fatal rate.

A numerical solution with python code has been found in Kaggle (Website 1. , 2020).

Note that, both of this model estimate rates that are not time dependent. There are other modifications of these type of models discussed in section 2.

4 Materials

For this project, different data sources have been used. For the covid-19 infected, recovered and deaths information for each country by day has been taken from Zindi competition website (https://zindi.africa/competitions/predict-the-global-spread-of-covid-19/data). Country specific economic, health and demographic indicator information have been taken from the world bank data repository (https://databank.worldbank.org/home.aspx). Note that, I had to spent most of my time to prepare these data. These data contain about 1500 indicators for each country from which I have selected (manually) 32 indicators that are relevant for this analysis. There were lots of missing for which I had to change some of the data manually and with different google source. I also used some statistical techniques to impute some data. It was also a difficult part to join these data with Zindi competition data. Because the country names are not same, and I had to manually change country names. The list of indicators considered are given bellow.

Access to electricity (% of population)

EG.ELC.ACCS.ZS

Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)

SH.DTH.COMM.ZS

Current health expenditure per capita (current US$)

SH.XPD.CHEX.PC.CD

Domestic general government health expenditure per capita (current US$)

SH.XPD.GHED.PC.CD

Domestic private health expenditure per capita (current US$)

SH.XPD.PVTD.PC.CD

GDP per capita (current US$)

NY.GDP.PCAP.CD

GNI per capita, Atlas method (current US$)

NY.GNP.PCAP.CD

Immunization, DPT (% of children ages 12-23 months)

SH.IMM.IDPT

Immunization, HepB3 (% of one-year-old children)

SH.IMM.HEPB

Immunization, measles (% of children ages 12-23 months)

SH.IMM.MEAS

Incidence of tuberculosis (per 100,000 people)

SH.TBS.INCD

International migrant stock (% of population)

SM.POP.TOTL.ZS

International tourism, number of arrivals

ST.INT.ARVL

Land area (sq. km)

AG.LND.TOTL.K2

Life expectancy at birth, total (years)

SP.DYN.LE00.IN

Merchandise exports by the reporting economy (current US$)

TX.VAL.MRCH.WL.CD

Merchandise exports to high-income economies (% of total merchandise exports)

TX.VAL.MRCH.HI.ZS

Merchandise imports from high-income economies (% of total merchandise imports)

TM.VAL.MRCH.HI.ZS

Merchandise trade (% of GDP)

TG.VAL.TOTL.GD.ZS

Net migration

SM.POP.NETM

Net primary income (Net income from abroad) (current US$)

NY.GSR.NFCY.CD

Population ages 0-14 (% of total population)

SP.POP.0014.TO.ZS

Population ages 15-64 (% of total population)

SP.POP.1564.TO.ZS

Population ages 65 and above (% of total population)

SP.POP.65UP.TO.ZS

Population density (people per sq. km of land area)

EN.POP.DNST

Population, total

SP.POP.TOTL

Rural population (% of total population)

SP.RUR.TOTL.ZS

Trade in services (% of GDP)

BG.GSR.NFSV.GD.ZS

Tuberculosis case detection rate (%, all forms)

SH.TBS.DTEC.ZS

Tuberculosis treatment success rate (% of new cases)

SH.TBS.CURE.ZS

UHC service coverage index

SH.UHC.SRVS.CV.XD

5 Methods

Much effort has been put for the data preparation and feature generation. Here is the diagram that will show the steps and methods has been considered for data preparation and modeling.

- Calculate growth distribution with respect to time for all countries and using k-means cluster method find group of country showing similar growth and use this cluster assignment as a feature
- For each country using Infected, Deaths, Recovered info fit SIR & SIRD model and predict susceptible, infected, recovered, fatal of all-time sequence and use these as a features for second step model
- Using Infected, Deaths, Recovered, country specific indicators, and some newly created features fit k-means model to get similar structure and condition in the same group and use it as a feature
- Find four closest neighbor countries and use those countries information as a new feature for each country. The logic behind this is that, if a neighboring country is in a very bad condition, there is a high probability that it may affect other neighbor country. For example, in Europe, the neighboring countries of Italy is also highly affected.
- Create some features based on country specific infected, deaths, recovered information. For example, how many days it took to have 100 infected, deaths or recovered people.
- Using time series change point trend model, find the change point where it started to grow or fall down the curve. Using that model predict for all time points and use these as features for next step model.
- Above steps have been used for all 173 countries and combined and merge these data to fit further machine learning model.
- To fit the machine learning models, I have considered Support vector regression, XGBoost regression, spline regression, random forest regression and neural network regression.
- Note that, it was found that spline regression has better performance for countries with high death and infected rates. That is way, I have considered boosting different models using spline regression. I have considered different (three) random subset of features (like random forest) to find different spline regression estimate and prediction and each of these predictions has been used in the final ensemble model.
- Finally using ensemble model (simple average), find the final daily prediction for each country.

For all the model, grid search method has been used to tune the parameters. Note that there are lots of free parameters, it was not possible to do extensive search for best parameters though I wanted to consider Bayesian framework or gradient descent hyperparameter tuning but could not implement it.

Here is the list of parameters I have considered for tuning.

Random forest:

```
'n_estimators': [30,  50, 80, 100, 150],
'max_depth':[3, 4,5,6],
'min_samples_leaf': [15, 20, 25, 30 , 50]
```

SVR:

```
'kernel':('poly', 'rbf', 'sigmoid'),
'C':np.linspace(1e-8,1, 30),
'epsilon': np.linspace(1e-8,.2, 10)
```

Spline Regression:

```
'max_degree': [1,2],
'penalty':np.linspace(1e-8, 50, 100)
```

XGBoost Regression:

```
'n_estimators': [50, 80, 100, 120, 150, 200],
'max_depth': [3,4,5,6,7],
'learning_rate': [1e-8],
'gamma': np.linspace(1e-8, 5, 25),
'subsample': np.linspace(.1, 1, 10),
'reg_alpha': [1e-8]
```

Neural Network:

```
'hidden_layer_sizes': [(5,), (10,), (20,), (50,), (10,5),

                    (20,10),(30,15), (20,10,5), (30,15,10)],
'alpha': np.linspace(1e-8, 1, 5),
'beta_1': np.linspace(1e-8, .99999, 5),
'beta_2': np.linspace(1e-8, .99999, 5),
'epsilon': np.linspace(1e-8, 1, 5),
'n_iter_no_change': [10]
```

6 Results

For this analysis, 174 countries have been considered. Data contains daily deaths, infected and recovered for each country from 22 January 2020 to 21 April 2020. All dates bellow 08 April 2020 has been considered as the training period and from 08 April 2020 to 21 April 2020 data as the test period. Table 2: Mean Absolute Error of observed and estimated cumulative deaths for selected countries. shows that support vector regression and XGBoost regression model performs best among these models. Mean absolute error (MAE) for test data of support vector regression (SVR), XGBoost regression and random forest regression (RFR) are even lower than the training MAE. However, it is seen that the XGboost and SVR models do not perform well for countries with highly infected and deaths rates (Figure 1, Figure 2, Table 2). It can be seen form Figure 1 that mean absolute error for XGBoost and SVR models are very high. These models underestimate the response for both training and testing data though all models underestimate the testing period. Figure 2 shows the observed and estimated cumulative deaths for USA (See

Appendices for more plots of different countries). Figure 2 shows that XGBoost and SVR models are highly underestimated the cumulative deaths for USA as well as countries with highly infected and deaths rates. On the other hand, spline regression (SPR) performs worst on the overall data but performs best for countries with highly infected and deaths rates (Table 2, Figure 2). It suggests that if we can fit some hierarchical models or if we can include some interaction and higher order terms in the model, we might get better performance from these models.

Table 1: Test mean absolute error for different models at hyperparameter tuning

Model	Test Mean Absolute Error	
	Train	Test
SVR	383.75	314.60
XGBoost	391.97	322.14
Random Forest	732.89	722.76
Spline Regression (All Variable)	2314.61	496.89
Spline Regression (Subset 1)	1631.72	1578.33
Spline Regression (Subset 2)	6493.71	8634.68
Neural Network	2172.07	2314.88
Ensemble	1715.86	2028.80

Table 2: Mean Absolute Error of observed and estimated cumulative deaths for selected countries.

Model	Mean Absolute Error (Final model)							
	USA		Italy		China		Bangladesh	
	Train	Test	Train	Test	Train	Test	Train	Test
SVR	1007.42	28995.93	2898.43	21278.93	2195.99	3804.57	1.47	56.29
XGBoost	1000.27	28953.62	2886.70	21236.64	2176.48	3762.30	18.03	24.11
RF	374.90	19250.55	1247.86	4760.96	723.58	793.61	1.23	42.37
SPR (All)	83.79	9314.05	972.75	3496.74	344.06	496.89	18.24	170.19
SPR(Set1)	87.11	11638.08	902.37	120.00	331.04	467.74	5.16	28.38
SPR(Set2)	93.20	1128.31	270.93	6170.33	146.14	2667.25	1.89	24.24
NN	143.72	10459.02	395.64	1247.01	336.57	1080.88	4.20	26.30
Ensemble	120.10	10307.36	657.08	1728.33	217.46	245.79	3.48	23.95

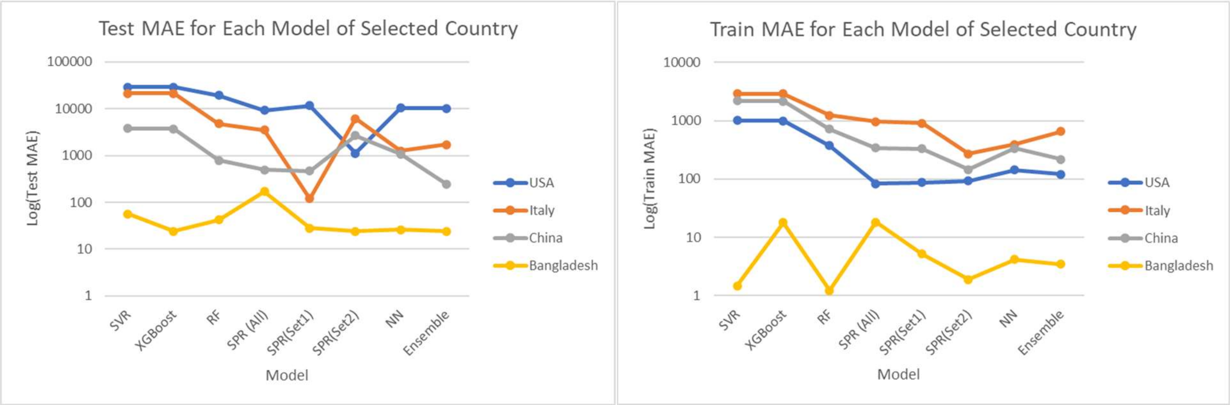
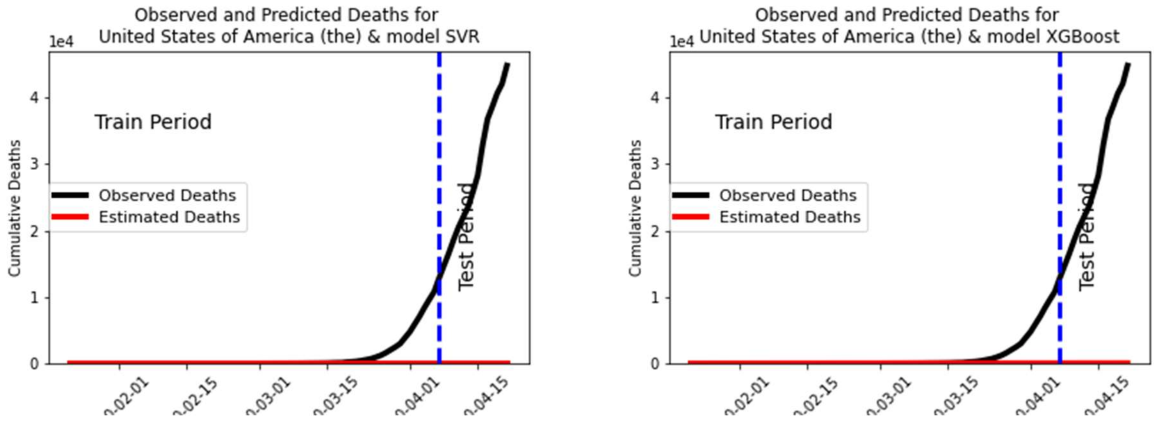


Figure 1: Train and Test MAE for different model of selected countries.



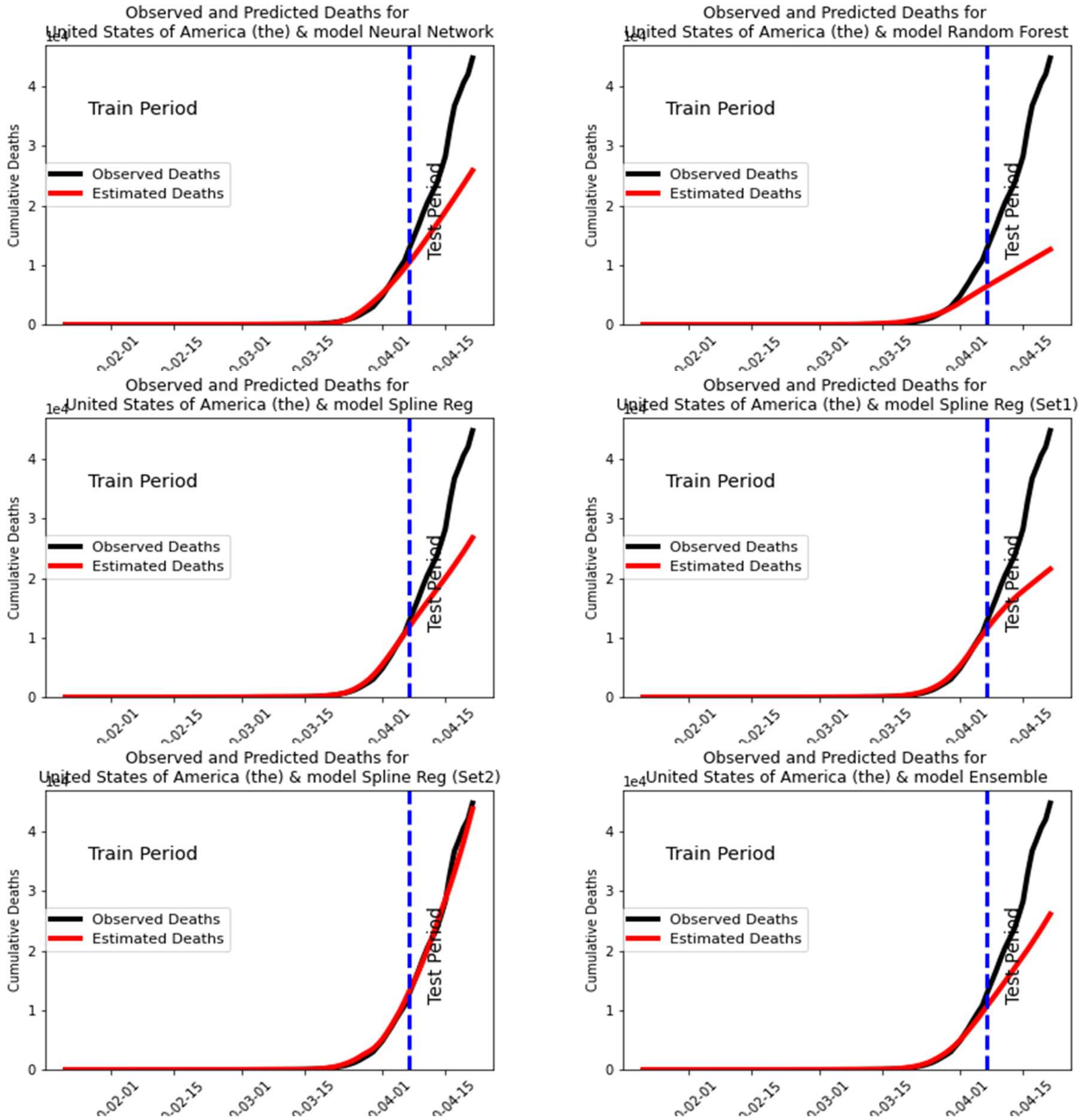


Figure 2: Observed and estimated cumulative deaths of USA for different model, (1) SVR, (2) XGBoost, (3) Ensemble, (4) Neural Network, (5) Random Forest, (6) Spline Reg. (all), (7) Spline Reg. (set1), (8) Spline Reg. (set2).

7 Conclusions

It is found that XGBoost model performs better for countries with lower infected and deaths rates. Spline regression model performs better for countries with higher infected and deaths rates though it performs poorly for other countries. It is found that all model underestimates for the longer time of the test data.

7.1 Limitation and future work

All machine learning model considers that the data distribution in the training window and test window will be same. However, for this coved-19 prediction problem, we are using different distribution for training and testing window. It is well known that for any type of epidemic, the distribution of infected or deaths will follow some sort of symmetric distribution. That is, from the beginning of the spread, the curve will go up until a certain time and then it will start to decline. So, the time window I have used for training model is the half part of the curve and predicting the other half part of the curve. That is way, for the long run testing periods, these models are underestimating that is prediction goes up.

Instead of predicting deaths directly if we can predict growth rate over the time and can select a model that can consider the functional form of such curve (like change point model), it would be easier to predict unknown part of the curve. For example, if we can just predict the pick point (time) of the curve and increasing rates over the time until that pick point, we can easily predict the down slop of the curve.

Additionally, in all those models, higher order polynomial and interaction terms has not been considered. It might be possible to have a better prediction if we could use hierarchical linear and nonlinear model. For example, if we could you a hierarchical model that will consider different slope for different time window based on the change point of the curve.

Another problem is that none of these models have any restriction on the output support. That is, we are predicting number of deaths and it is always greater than 0. However, for all these models I have considered that the response is either on the real line or (like spline regression) response is coming from normal distribution that is also on real line. If we can consider the same models on restricted support (e.g. Poisson distribution which represents the count observations), would be more appropriate.

8 Acknowledgement

This work is a solution to Zindi competition: Predict the Global Spread of COVID-19 (https://zindi.africa/competitions/predict-the-global-spread-of-covid-19). Thanks to my friends Sandip Sureka and Shiplu Sarker to help me identify country indicators and to understand SIR and SIRD model.

9 Bibliography

Amaro, J. E. (2020). The D model for deaths by COVID-19. *arXiv preprint*, 2003.13747.

Bayes, C., Rosas, G. G., & Valdivieso, L. (2020). Modelling death rates due to COVID-19: A Bayesian approach. *arXiv*. Retrieved from arXiv: https://arxiv.org/pdf/2004.02386.pdf

Calafiore, G. C., Novara, C., & Possieri, C. (2020). A Modified SIR Model for the COVID-19 Contagion in Italy. *arXiv preprint*, 2003.14391.

Chen, Y.-C., Lu, P.-E., & Liu, T.-H. (2020). A Time-dependent SIR model for COVID-19 with Undetectable Infected Persons. *arXiv preprint*, 2003.00122.

Dhanwant, J. N., & Ramanathan, V. (2020). Forecasting COVID 19 growth in India using Susceptible-Infected-Recovered. *arxiv*. Retrieved from https://arxiv.org/ftp/arxiv/papers/2004/2004.00696.pdf

Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons and Fractals, 134*.

Harko, T., Lobo, F. S., & Mak, M. K. (2014). Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. *Applied Mathematics and Computation*, 184-194. Retrieved from https://arxiv.org/pdf/1403.2160.pdf

IHME Team, C.-1. h. (2020). Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries. *medRxiv*.

Johnson, T. (n.d.). Mathematical Modeling of Diseases: Susceptible-Infected-Recovered (SIR) Model. Retrieved from https://op12no2.me/stuff/tjsir.pdf

Keeling, M. J., & Rohani, P. (2008). *Modeling Infectious Diseases IN HUMANS AND ANIMALS.* New Jersey : Princeton University Press.

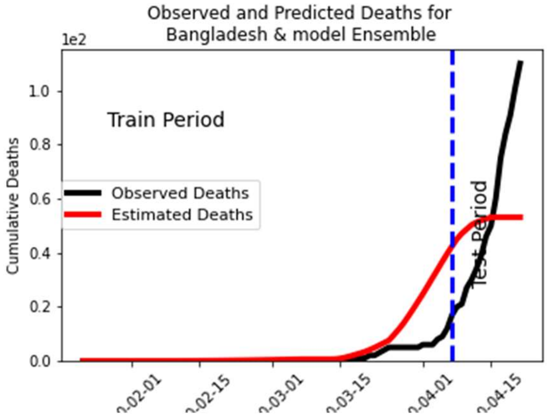
Kermack, W., & McKendrick, A. (1927). Contribution to the mathematical theory of epidemics. *Proc. Roy. Soc. Lond A, 115*, 700-721.

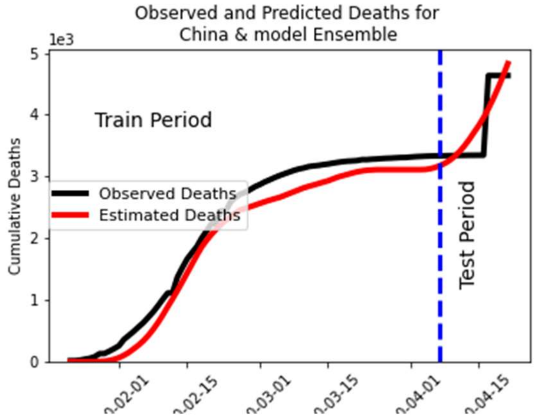
Qi, C., Karlsson, D., Sallmen, K., & Wyss, R. (2020). Model studies on the COVID-19 pandemic in Sweden. *arXiv preprint*, 2004.01575. Retrieved from https://arxiv.org/pdf/2004.01575.pdf

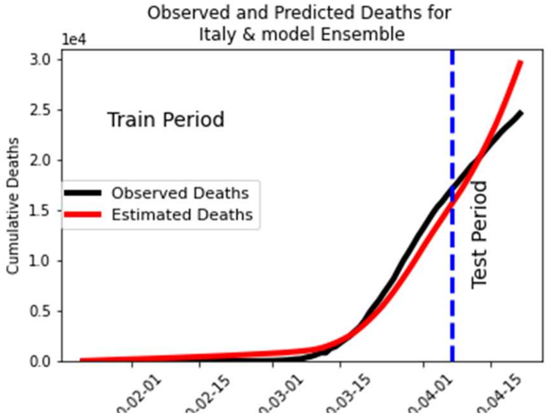
Website, 1. (2020). *COVID-19 data with SIR model*. Retrieved from https://www.kaggle.com/kstatucf/covid-19-data-with-sir-model/edit?rvi=1

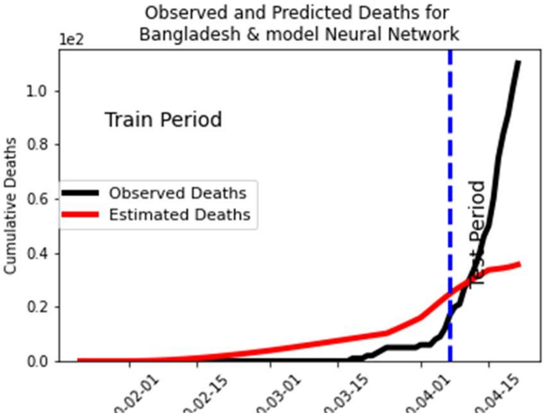
Website, 2. (2019). *Fitting the SIR model of disease to data in Python*. Retrieved from https://numbersandshapes.net/post/fitting_sir_to_data_in_python/?fbclid=IwAR0uHK4iijwmT6nw_Em3YGrUukhXS-Ho-qwDrIvU4wcGgcDwt9vo6JRScdU

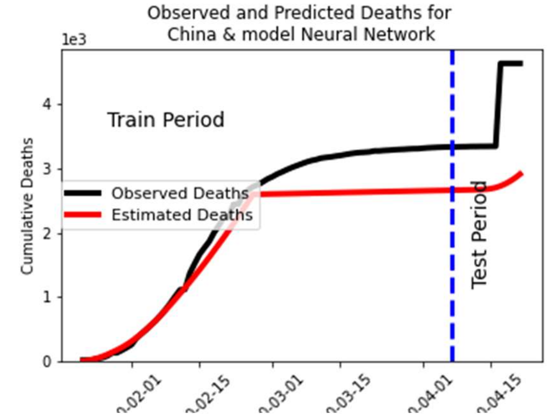
10 Appendices

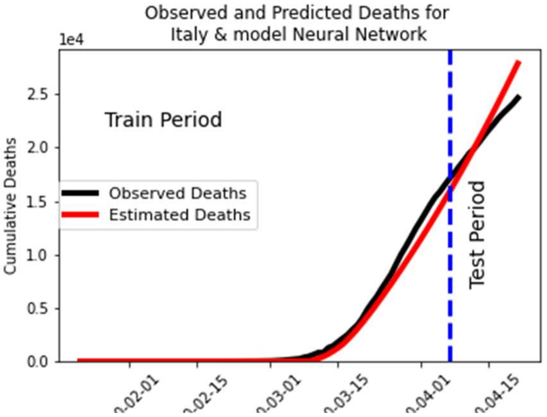

Observed and Predicted Deaths for Bangladesh & model Ensemble


Observed and Predicted Deaths for China & model Ensemble

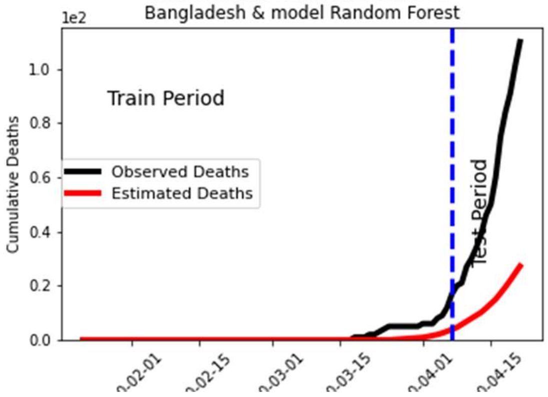

Observed and Predicted Deaths for Italy & model Ensemble


Observed and Predicted Deaths for Bangladesh & model Neural Network

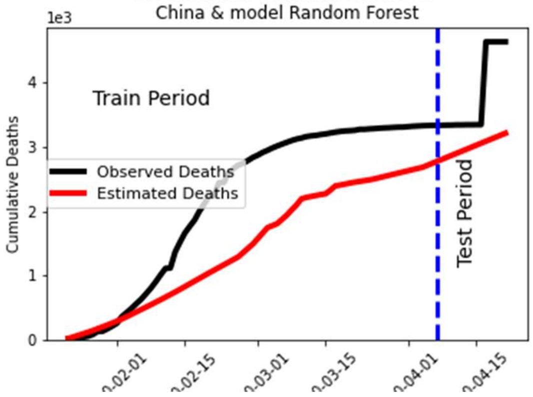

Observed and Predicted Deaths for China & model Neural Network


Observed and Predicted Deaths for Italy & model Neural Network

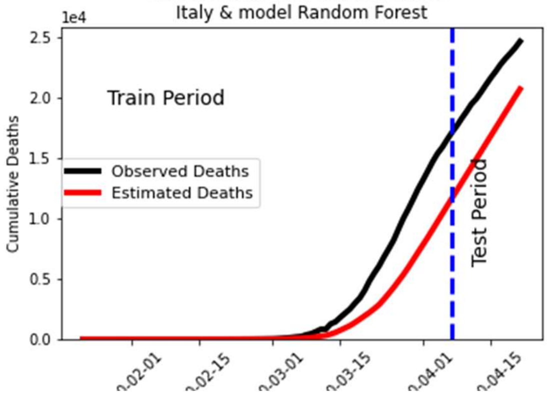
Observed and Predicted Deaths for
Bangladesh & model Random Forest



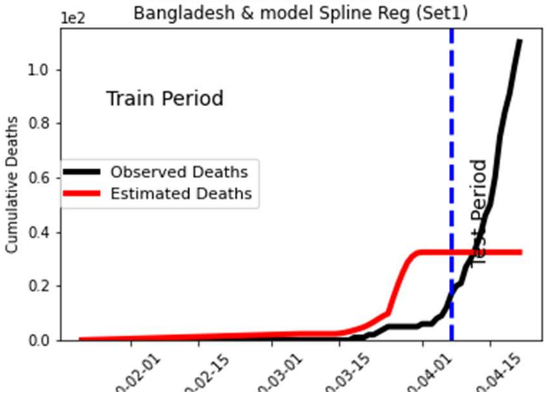
Observed and Predicted Deaths for
China & model Random Forest



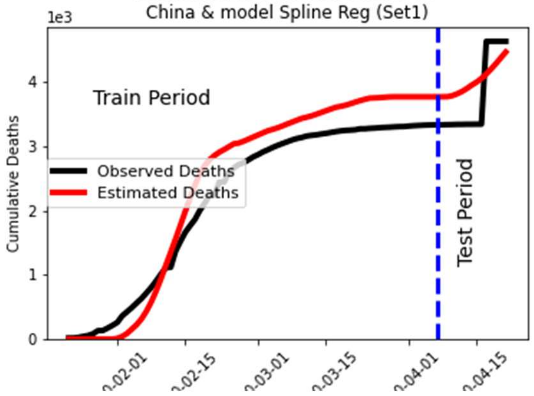
Observed and Predicted Deaths for
Italy & model Random Forest



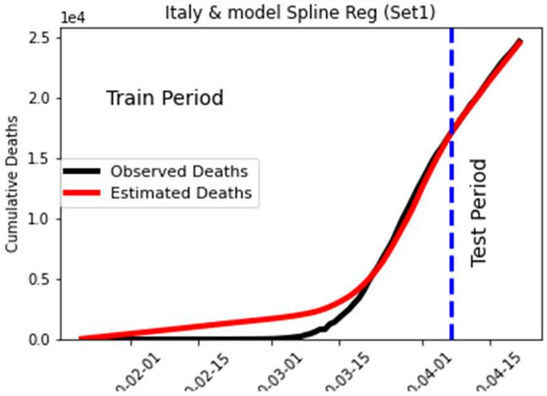
Observed and Predicted Deaths for
Bangladesh & model Spline Reg (Set1)



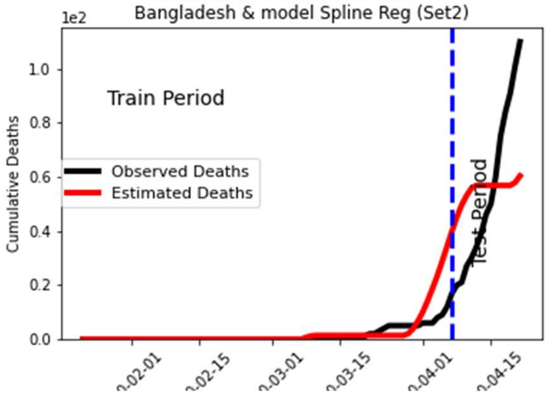
Observed and Predicted Deaths for
China & model Spline Reg (Set1)



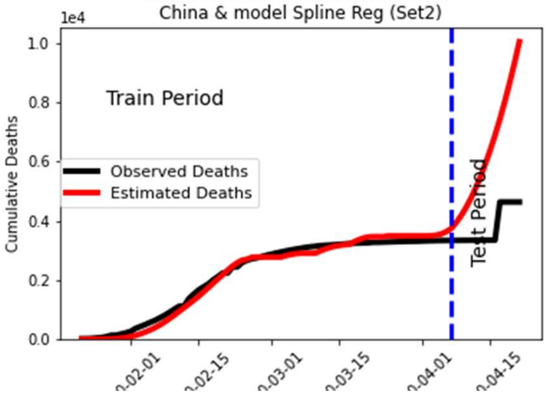
Observed and Predicted Deaths for
Italy & model Spline Reg (Set1)

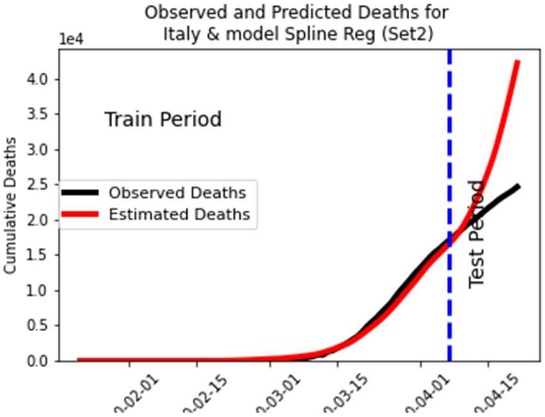


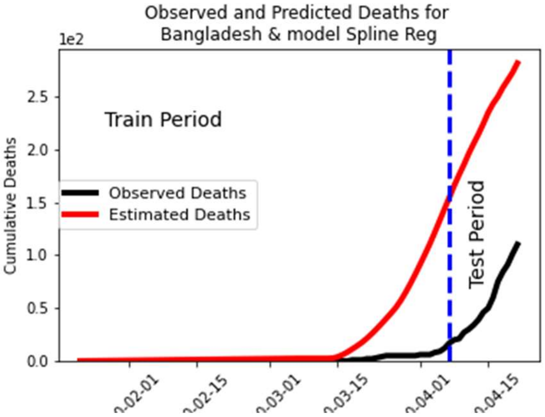
Observed and Predicted Deaths for
Bangladesh & model Spline Reg (Set2)

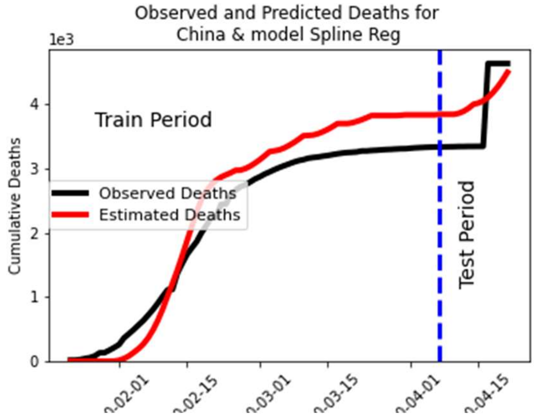


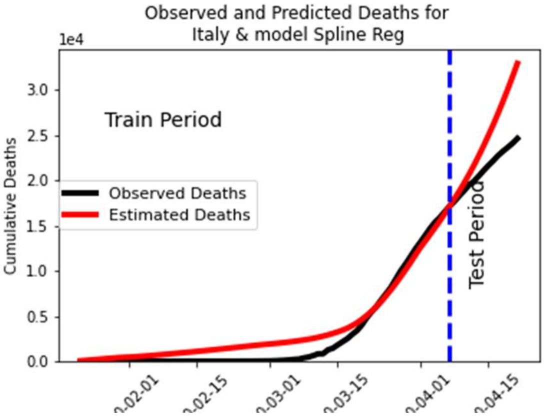
Observed and Predicted Deaths for
China & model Spline Reg (Set2)

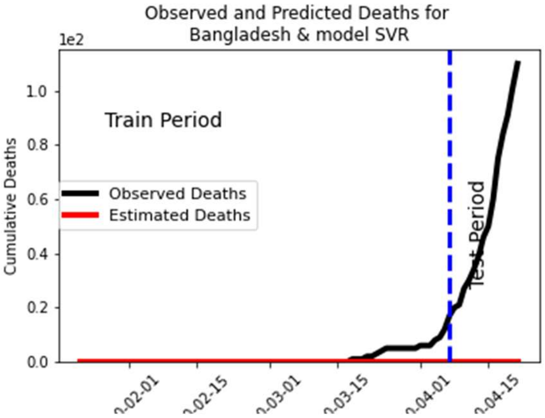


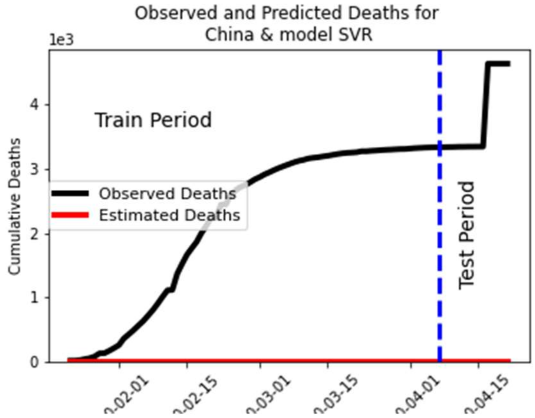

Observed and Predicted Deaths for
Italy & model Spline Reg (Set2)
1e4
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Cumulative Deaths
Train Period
Observed Deaths
Estimated Deaths
Test Period
0-02-01
0-02-15
0-03-01
0-03-15
0-04-01
0-04-15

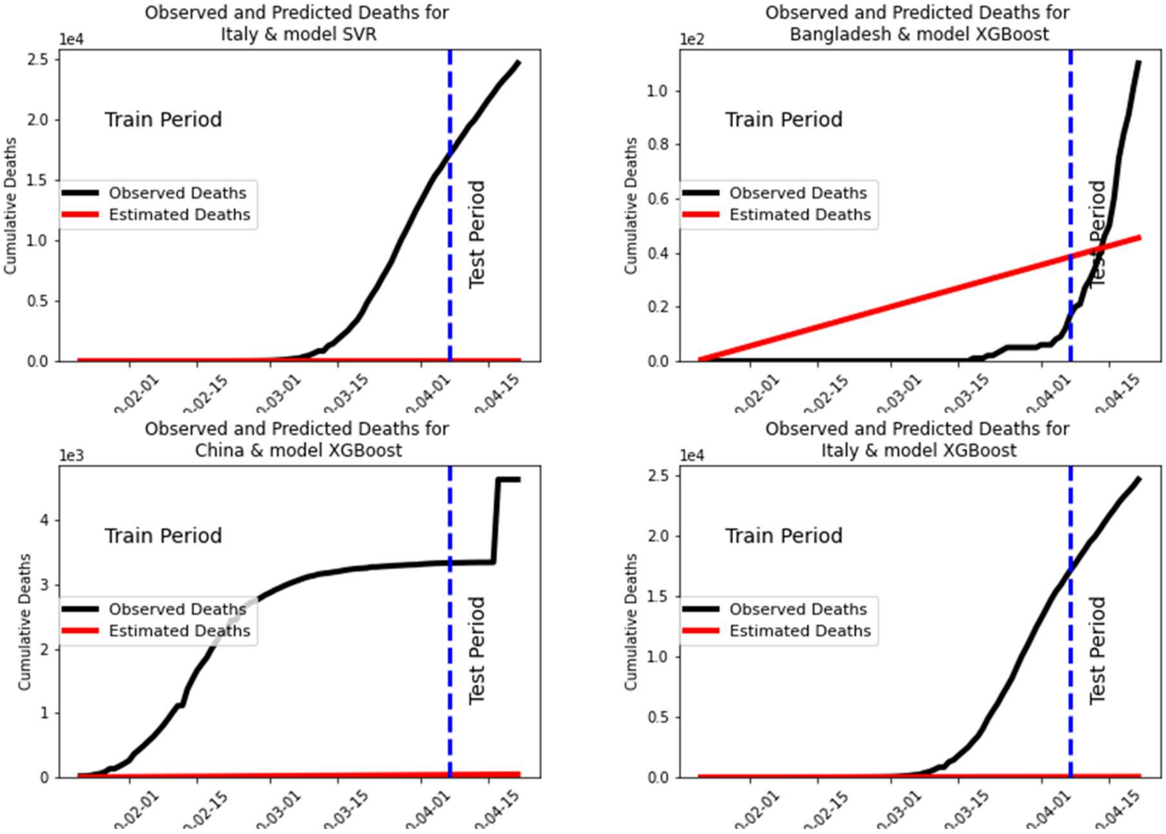

Observed and Predicted Deaths for
Bangladesh & model Spline Reg
1e2
2.5
2.0
1.5
1.0
0.5
0.0
Cumulative Deaths
Train Period
Observed Deaths
Estimated Deaths
Test Period
0-02-01
0-02-15
0-03-01
0-03-15
0-04-01
0-04-15


Observed and Predicted Deaths for
China & model Spline Reg
1e3
4
3
2
1
0
Cumulative Deaths
Train Period
Observed Deaths
Estimated Deaths
Test Period
0-02-01
0-02-15
0-03-01
0-03-15
0-04-01
0-04-15


Observed and Predicted Deaths for
Italy & model Spline Reg
1e4
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Cumulative Deaths
Train Period
Observed Deaths
Estimated Deaths
Test Period
0-02-01
0-02-15
0-03-01
0-03-15
0-04-01
0-04-15


Observed and Predicted Deaths for
Bangladesh & model SVR
1e2
1.0
0.8
0.6
0.4
0.2
0.0
Cumulative Deaths
Train Period
Observed Deaths
Estimated Deaths
Test Period
0-02-01
0-02-15
0-03-01
0-03-15
0-04-01
0-04-15


Observed and Predicted Deaths for
China & model SVR
1e3
4
3
2
1
0
Cumulative Deaths
Train Period
Observed Deaths
Estimated Deaths
Test Period
0-02-01
0-02-15
0-03-01
0-03-15
0-04-01
0-04-15


Observed and Predicted Deaths for
Italy & model SVR
1e4
2.5
2.0
1.5
1.0
0.5
0.0
Cumulative Deaths
Train Period
Test Period
Observed Deaths
Estimated Deaths
0-02-01
0-02-15
0-03-01
0-03-15
0-04-01
0-04-15
Observed and Predicted Deaths for
Bangladesh & model XGBoost
1e2
1.0
0.8
0.6
0.4
0.2
0.0
Cumulative Deaths
Train Period
Test Period
Observed Deaths
Estimated Deaths
0-02-01
0-02-15
0-03-01
0-03-15
0-04-01
0-04-15
Observed and Predicted Deaths for
China & model XGBoost
1e3
4
3
2
1
0
Cumulative Deaths
Train Period
Test Period
Observed Deaths
Estimated Deaths
0-02-01
0-02-15
0-03-01
0-03-15
0-04-01
0-04-15
Observed and Predicted Deaths for
Italy & model XGBoost
1e4
2.5
2.0
1.5
1.0
0.5
0.0
Cumulative Deaths
Train Period
Test Period
Observed Deaths
Estimated Deaths
0-02-01
0-02-15
0-03-01
0-03-15
0-04-01
0-04-15