COVID-19 US Cases Projection Using Korea and Italy as Leading Indicators

Qiuping Xu
qx0731@gmail.com
Seattle, WA

Introduction

As Coronavirus progresses in US, we want to detect the lagging effect and predict what will happen next in US through the analysis of Korea, Italy and US data. In the article, we use the Kaggle open source data and employed sample cross correlation function (CCF) to identify the lagging between US and Korea, US and Italy. In the end, we applied DoD change on the registered curves to provide an estimate of US COVID-19 case in the next week under different scenarios. On 03/18, based on the data evidence, we added in exponential model.

Data and Code are open sourced at https://github.com/qx0731/covid_19_projection_03_16. With daily updated data, users could run a daily refresh with minimal effort.

US Projections on 03/17

Table 1: Most recent projection of US COVID-19 cases. On 03/18, besides two scenarios, we also added the exponential increase prediction.

Date	Based on Korea	Based on Italy	Exp
3/3/20	118	118	118
3/4/20	176	176	176
3/5/20	223	223	223
3/6/20	341	341	341
3/7/20	417	417	417
3/8/20	584	584	584
3/9/20	778	778	778
3/10/20	1,053	1,053	1,053
3/11/20	1,315	1,315	1,315
3/12/20	1,922	1,922	1,922
3/13/20	2,450	2,450	2,450
3/14/20	3,173	3,173	3,173
3/15/20	4,019	4,019	4,019
3/16/20	5,723	5,723	5,723
3/17/20	7,731	7,731	7,731
3/18/20	8,367	8,555	10,408
3/19/20	9,118	10,504	13,938
3/20/20	9,819	12,739	18,665
3/21/20	10,352	14,885	24,995
3/22/20	10,711	17,833	33,471
3/23/20	10,901	20,859	44,822
3/24/20	11,253	23,584	60,023
3/25/20	11,418	26,556	80,379
3/26/20		30,102	107,639

Data and Analysis

Date:

The data are all from Kaggle or from GitHub

- Korea data: https://www.kaggle.com/kimjihoo/coronavirusdataset
- Italy data: https://www.kaggle.com/sudalairajkumar/covid19-in-italy

- USA data: https://www.kaggle.com/sudalairajkumar/covid19-in-usa

Visualization:

From those raw datasets, we extracted the daily update on the cumulative total of test cases, positive cases, released cases and decreased cases. The three plots in appendix show the time series of those four tracking metrics in the 3 countries. However, those plots are not easy to digest or to register. To remove the effect of the scale(population) and the temporal effort, we calculated the DoD change on the cumulative positive cases [see Fig.1.]. To better visualize the plots, we scale the time periods all to 60 days. The upper left is for Korea from date 2020-01-21 to 2020-03-12, the upper right is for Italy from date 2020-02-24 to 2020-03-16, while the bottom one is for US from 2020-03-03 to 2020-03-15. The intuitive explanation is that when the DoD is bigger than 1, we will see an increase in the total confirmed cases, while DoD is close to 1, with the released cases, we will see a decrease in the total remaining cases.

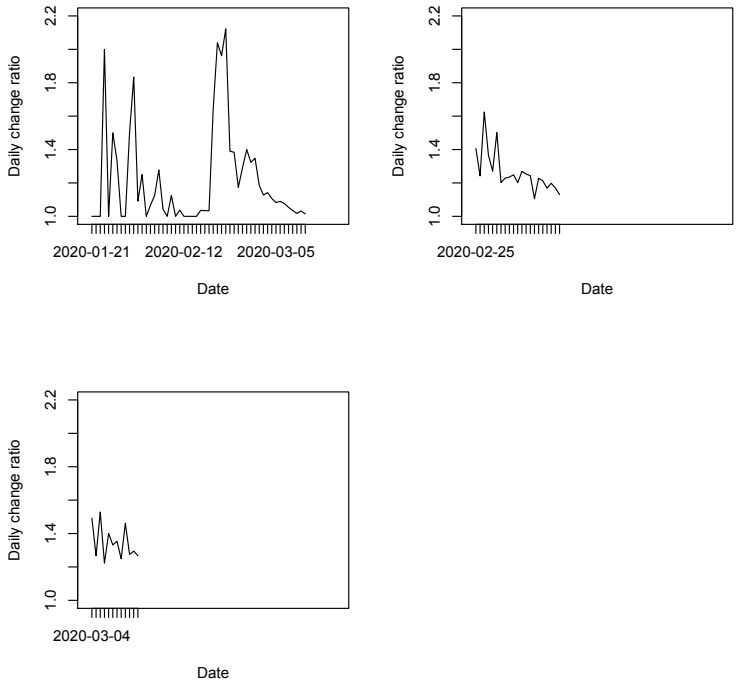


Figure 1: DoD (%) change of total confirmed cases in Korea (upper left), Italy (upper right) and US (bottom).

Through the visual inspection, we can see the Italy curve and US curve look similar. We used sample cross correlation function (CCF) to identify the lagging of between those two curves and confirmed with our hypothesis. The correlation coefficient with lagging 0 is the largest with value at 0.51 [See Fig.2.]. Given the date difference in the dataset, we can say 2020-02-25 data in Italy is similar to the 2020-03-04 data in US (**Italy is leading US by around 10 days**).

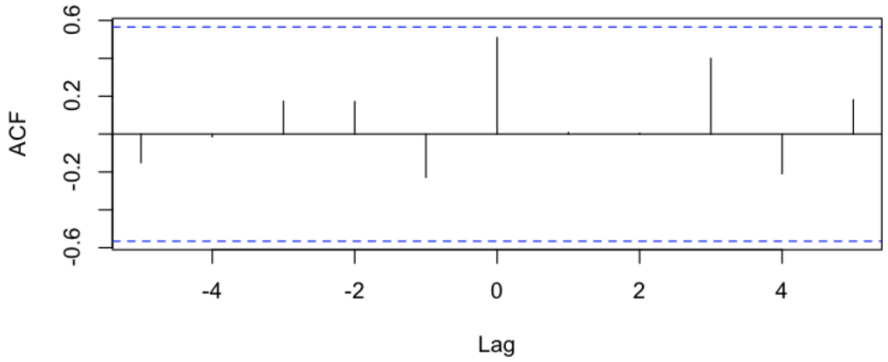
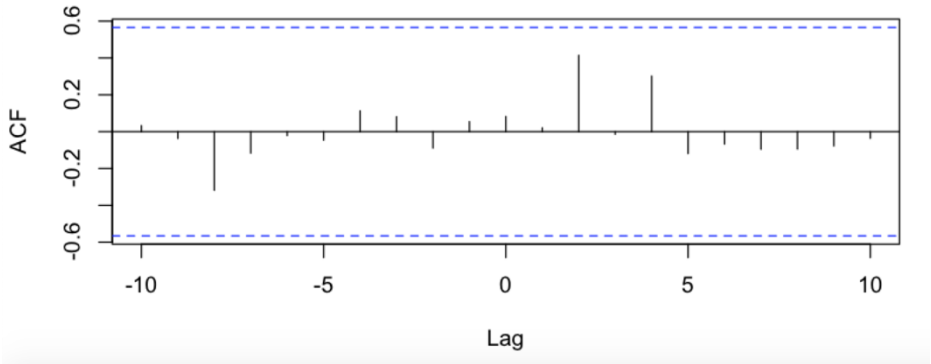


Figure 2: CCF between DoD change in Italy and US.

For Korean data, we restrict the data from 2020-02-18 onward and run the same CCF. We can say 2020-02-20 data in Korea is similar to the 2020-03-04 data in US (**Korea is leading US by around 15 days**).



<u>Projections:</u>

Based on the learned lagging, we tried to run the projection use Korean and Italian's DoD data [See Fig.3]. On 03/17, based on the data evidence [table2], we added in the projection of using log-linear regression between the log(case number) and days.

Table 2: The model summary of the log-linear model between log(case) and days (1...n). ***Sadly,*** *this model fits the current data very well (with adjusted R^2 0.998).*

```
> summary(crazy_model)

Call:
lm(formula = log_case ~ days, data = us_time_case)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09931 -0.02220 -0.00500  0.03719  0.08582

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.577968   0.028359  161.43   <2e-16 ***
days        0.292024   0.003119   93.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05219 on 13 degrees of freedom
Multiple R-squared:  0.9985,    Adjusted R-squared:  0.9984
F-statistic:  8766 on 1 and 13 DF,  p-value: < 2.2e-16
```

Next

More analysis can be done on the test capacity, death rate and relationship among all the metrics.

Reference

1. Korea data: https://www.kaggle.com/kimjihoo/coronavirusdataset
2. Italy data: https://www.kaggle.com/sudalairajkumar/covid19-in-italy
3. USA data: https://www.kaggle.com/sudalairajkumar/covid19-in-usa
4. Sample cross correlation function (CCF): https://nwfsc-timeseries.github.io/atsa-labs/sec-tslab-correlation-within-and-among-time-series.html

Date	Based on Korea	Based on Italy
3/3/20	118	118
3/4/20	176	176
3/5/20	223	223
3/6/20	341	341
3/7/20	417	417
3/8/20	584	584
3/9/20	778	778
3/10/20	1053	1053
3/11/20	1315	1315
3/12/20	1922	1922
3/13/20	2450	2450
3/14/20	3173	3173
3/15/20	4019	4019
3/16/20	4592	5038
3/17/20	5084	6266
3/18/20	5502	6933
3/19/20	5996	8513
3/20/20	6457	10325
3/21/20	6807	12065
3/22/20	7044	14454
3/23/20	7169	16906
3/24/20	7400	19115
3/25/20	7508	

Figure 3:Projects under different scenarios. I am hoping US will follow Korean trajectory, but based on 03/16 and 03/17 data, US is tracking closely with Italy ;(.

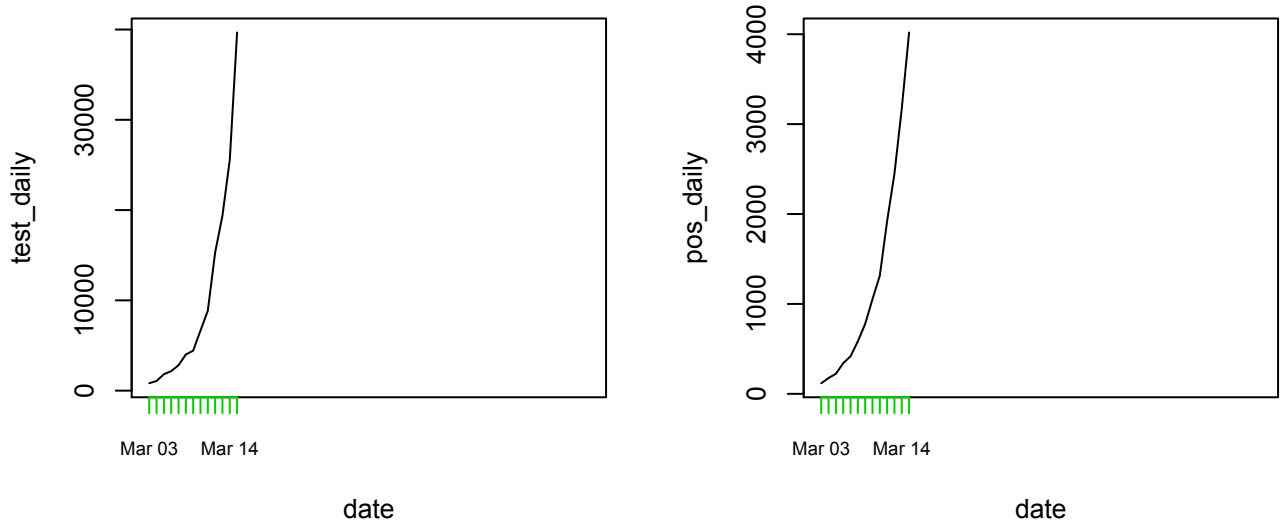
Date	Based on Korea	Based on Italy
3/3/20	118	118
3/4/20	176	176
3/5/20	223	223
3/6/20	341	341
3/7/20	417	417
3/8/20	584	584
3/9/20	778	778
3/10/20	1053	1053
3/11/20	1315	1315
3/12/20	1922	1922
3/13/20	2450	2450
3/14/20	3173	3173
3/15/20	4019	4019
3/16/20	5723	5723
3/17/20	6337	7117
3/18/20	6858	7876
3/19/20	7474	9671
3/20/20	8048	11728
3/21/20	8485	13704
3/22/20	8780	16418
3/23/20	8935	19204
3/24/20	9223	21712
3/25/20	9359	24449

Figure 4: Refresh with one day of new data.

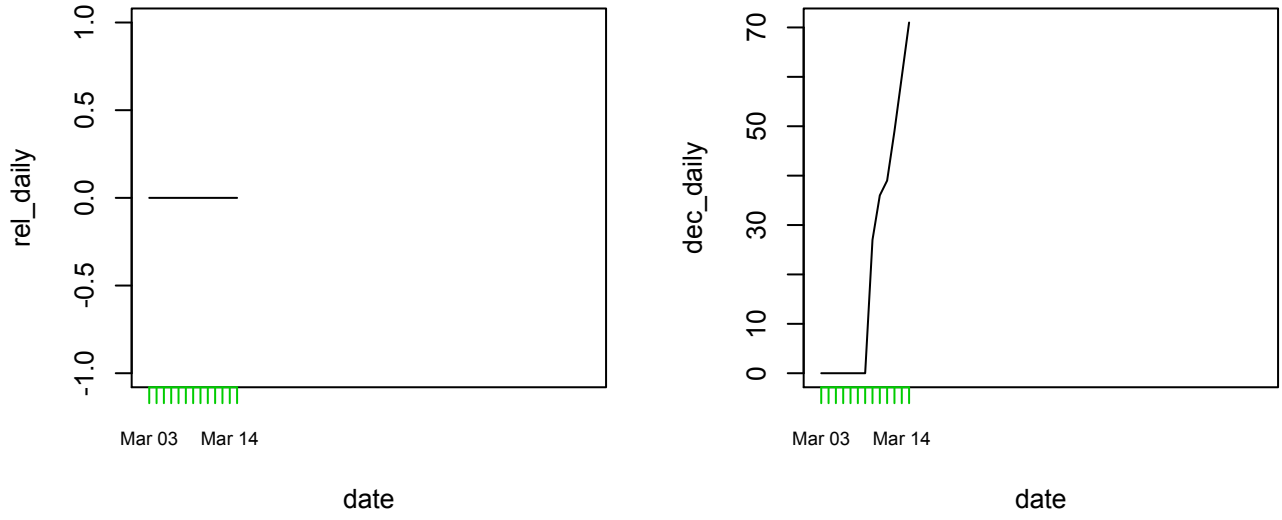
Date	Based on Korea	Based on Italy	Exp
3/3/20	118	118	118
3/4/20	176	176	176
3/5/20	223	223	223
3/6/20	341	341	341
3/7/20	417	417	417
3/8/20	584	584	584
3/9/20	778	778	778
3/10/20	1,053	1,053	1,053
3/11/20	1,315	1,315	1,315
3/12/20	1,922	1,922	1,922
3/13/20	2,450	2,450	2,450
3/14/20	3,173	3,173	3,173
3/15/20	4,019	4,019	4,019
3/16/20	5,723	5,723	5,723
3/17/20	7,731	7,731	7,731
3/18/20	8,367	8,555	10,408
3/19/20	9,118	10,504	13,938
3/20/20	9,819	12,739	18,665
3/21/20	10,352	14,885	24,995
3/22/20	10,711	17,833	33,471
3/23/20	10,901	20,859	44,822
3/24/20	11,253	23,584	60,023
3/25/20	11,418	26,556	80,379
3/26/20		30,102	107,639

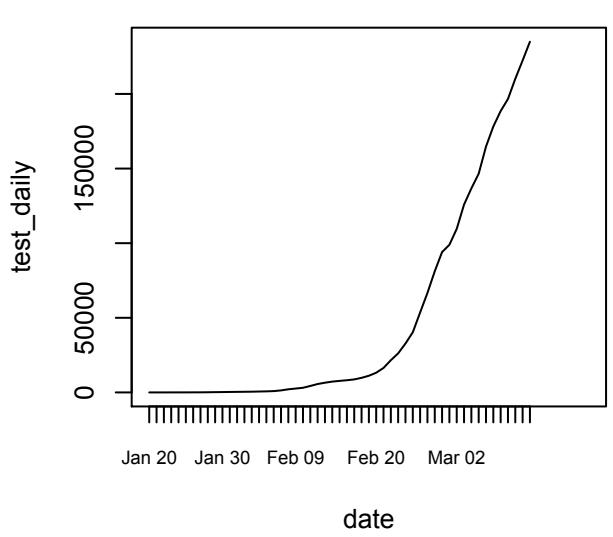
Figure 5: Refresh with one day of new data.

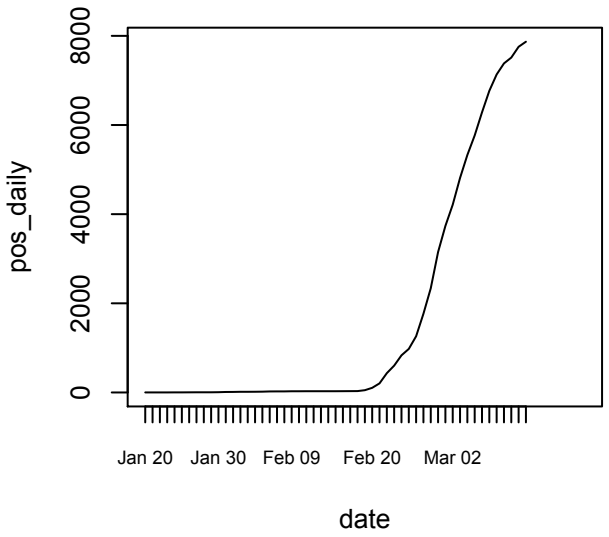
Appendix



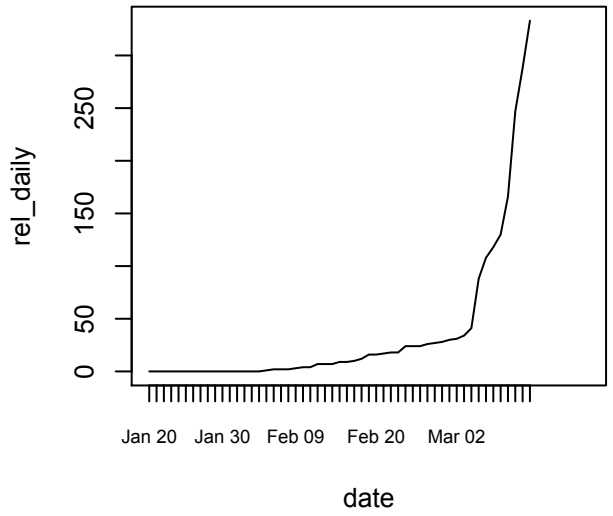
USA: Time Series(total tested, positive case, released, deceased) from 2020-03-04 to 2020-03-16

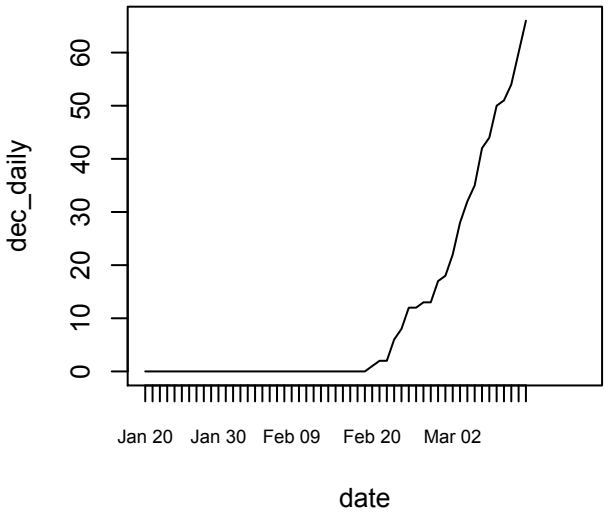




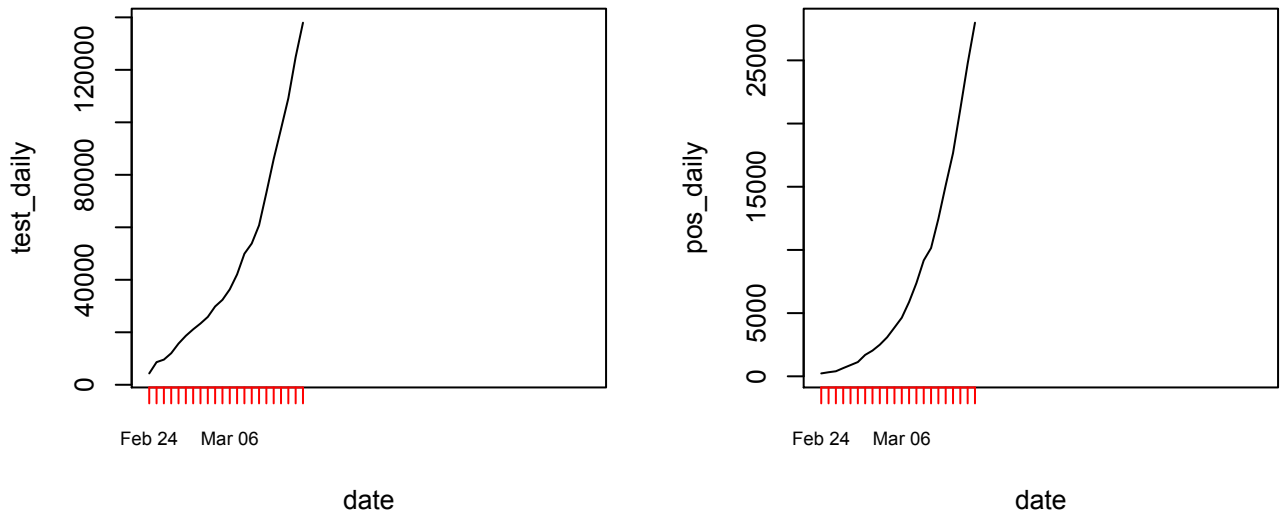


Korea: Time Series(total tested, positive case, released, deceased) from 2020-01-20 to 2020-03-12

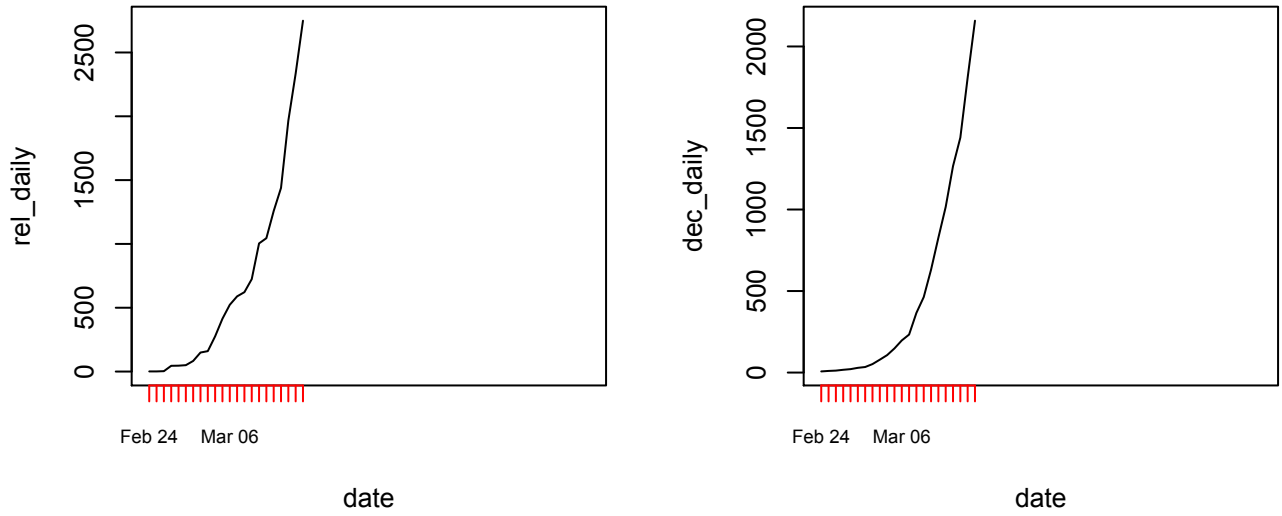




64



Italy: Time Series(total tested, positive case, released, deceased) from 2020-02-24 to 2020-03-16



65