Analysis of COVID-19 in South Korea and the United States

Riley Clarke, RileyWClarke, rwclarke@udel.edu, Visualization Manager

Jinbiao Ji, JinbiaoJi, jinbiao@udel.edu, Methodology manager

Adam Marrs, Marrs-jin, marrs@udel.edu, Communication/Literature Manager

Shaquann Seadrow, shaqsead16, sseadrow@udel.edu , Data Manager

Abstract

The handling of the Covid-19 outbreak in South Korea is unusual in how effective the country was in “flattening the curve” without imposing severe quarantine restrictions. We have done work trying to understand and characterize the outbreak in South Korea to gain insight into its success. Facebook Prophet’s modular regression model was chosen to analyze the death counts in Korea and look for changepoints. These changepoints were tentatively linked to policy decisions. Finally, we used this information to compare to several states in the United States. No states have shown the same success, though New York’s death curve is closest.

Introduction

The rapid spread of the Coronavirus (Covid-19) is an ongoing global issue responsible for widespread disruption. Our group will focus on the cases in South Korea, which started after January 20th, 2020. Presently, South Korea (SK) has had over 11000 confirmed cases of the virus and 262 deaths **[1]**. South Korea’s success can be attributed to a variety of factors: immediate cooperative action at the federal and local level, aggressive and early testing procedures, contact tracing, and citizen support **[2].** The results of these and more procedures have been a dramatic flattening in the number of cases confirmed, as well as a low death rate compared to other countries. **[3]**

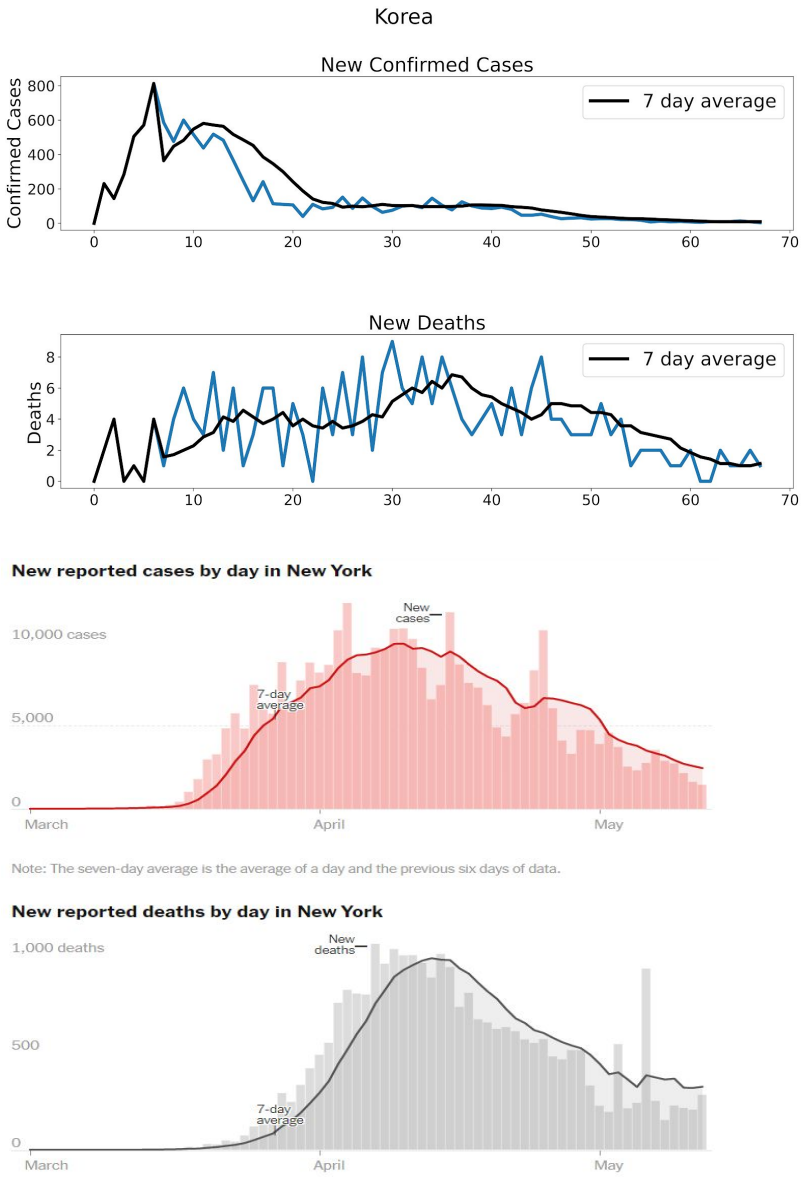

Figure 1: Graphs of the new reported deaths and cases for South Korea and New York. The data for South Korea starts at the day of 5 confirmed deaths, which is why it does not have the long tail at the beginning. The line in black is the average of the last 7 days for the count. Notice the similar shape of the two curves; this is not the case for the other states, who do not have strong declining tails at the right hand side. Also note the peaks of the graphs: New York's death by day peaks slightly after its reported cases by day, while Korea's death by day peaks 30 days after its peak in new cases. **[11]**

We have sought to understand specific things SK did to combat the virus, and if any of those actions have been/could be applied to the United States. We will look at a set of states, since the outbreak in the United States has not been uniform. New York is the most comparable, being densely populated and having enough time for the cases to start declining. California had

its first patient within a week of SK, Nebraska is a contrast to a less dense region, Florida had the unique case of Spring Break celebrations, and D.C. is another dense region and the capital.

When looking at changepoints, we combined data-specific points, like time since first death or the amount of lag between cases being confirmed and death, as well as societal change points. This encompasses things like social distancing measures, public noncompliance events, federal support, etc. While it is impossible to definitely link a data event with a policy event, we did find interesting patterns in the Korean data and the state data. We also used cross-correlation and distance metrics to see how close each state's death curve related to SK's.

Data

The South Korea data comes from DS4C (Data Science for Coronavirus), a dataset composed of the current pandemic figures from the KCDC (Korea Centers for Disease Control & Prevention) and available on Kaggle. DS4C is maintained by Project Manager, Joong Kun Lee, and collaborators. The data set provides the time series for cumulative testing (including positive and negative results), released patients, and deceased patients. DS4C also provides detail in locality, weather, and other epidemiological information that could be useful. The bulk of Covid19 in the United States are concentrated in metropolitan areas, such as New York and Chicago. Also, decisions regarding pandemic management have been made primarily the responsibility of the individual states. Recognizing these factors makes a comparison between the entire U.S. and South Korea infeasible, and we proceed with, for the U.S., individual state analyses. The data "Covid-19 in USA" provides covid19 time-series at the state level and is of equal utility to DS4C. It was created by SRK (the kaggle users name) composed of data from the COVID-19 Tracking Project and The NY Times. The outbreaks do not share a common start date, but we impose a cut off of April 30th, 2020 for all time series. We also use the "Covid-19 State Data" data set, because it contains the values of state population size and other variables that could be relevant to the pandemic.

Our analysis primarily utilizes the reported deaths more than positive cases. South Korea's more aggressive and larger -scale testing provides a more accurate understanding of the spread vs. the U.S. limited testing. Death's are less prone to testing bias. However, reporting is simultaneously ongoing with the pandemic, so we recognize the accuracy, in both counting and temporal, has limitations **[4]**. Examples of such complications include the frequency and delay to which deaths are being reported, especially overwhelmed localities, and possible undercounting due to the deaths of untested, infected individuals outside of medical facilities.

Dataset name	URL	Number of rows	Number of columns	Number of relevant columns	Number of valid rows	Data type
Data Science for Coronavirus (DS4C) **Files Used:** **1)time.csv**	https://www.kaggle.com/kimjihoo/coronavirusdataset	71	7	4	71	Datetime, int
Covid-19 in USA (actively updated) **Files Used:** **1)us_covid19_daily.csv** **2)us_states_covid19_daily.csv**	https://www.kaggle.com/sudalairajkumar/covid19-in-usa	**1) 32** **2) 1318**	**24** **25**	**4** **4**	**32** **1318**	Date, int
COVID-19 State Data **Files Used:** **COVID19_state.csv**	https://www.kaggle.com/nightranger77/covid19-state-data?select=COVID19_state.csv	**51**	**26**	**2**	**25**	Ordered by state. No date time

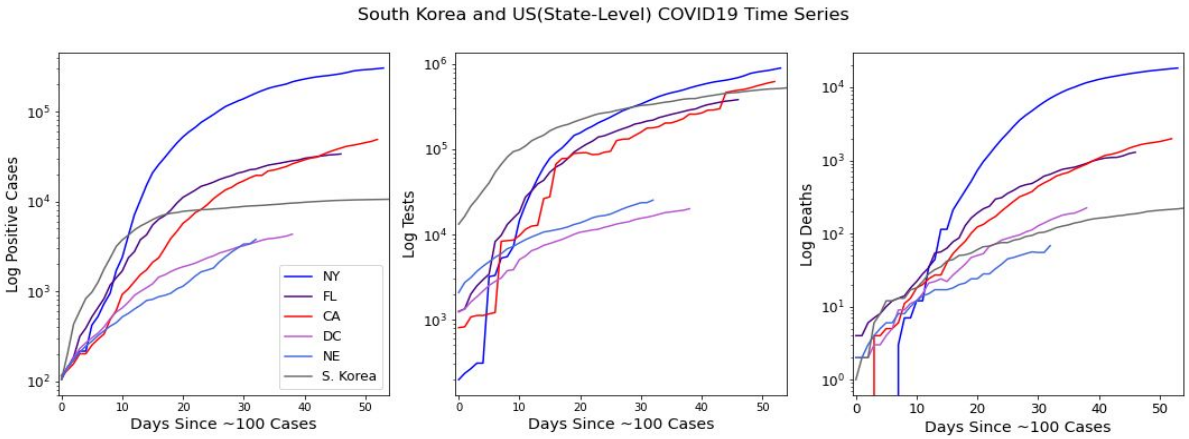

Figure 2: The visualization of the cumulative positive cases, tests and deaths with respect to time. All time series are initiated by the date of the first 100 or more cases.

The states in our analysis are California, Florida, Nebraska, New York, and Washington D.C. As discussed in the introduction, this set gives us diversity in locality, state type, and pandemic responses for our analyses. As you can see with positive cases and deaths S.K.'s curves flatten, while U.S. cases accelerate.

Data and Lags

Policy/events are not immediately linked to data and changepoints. For one, no policy is implemented immediately. There will be a delay between an event or policy and its implementation, as well as its widespread practice. Adding onto this, the majority of Covid-19 positive patients will remain asymptomatic for several days post-infection, and won't get tested until later. There will therefore be a lag (that is also influenced by region) between infection date and date of a "confirmed positive" data point. Confirmed death will have an even larger lag of being confirmed infected and dying, which we took into consideration. Using Korea and research **[13]** as a guideline, we expect any event or policy to have a 5-10 day delay in confirmed cases, and there to be a death lag unique to the region.

Other Methodologies

Epidemiologists have studied the spread of epidemics and successfully applied the classical logistic growth model to describe the data.**[8]** With the logistic type models, a research team was able to calibrate and model the reported number of confirmed cases in the outbreak of Covid-19 in China.**[9]** Logistic growth is characterized by increasing growth in the beginning period, but a decreasing growth at a later stage, as you get closer to a maximum. It's a nonlinear growth that saturates at a carrying capacity.

Logistic growth is defined by this formula:

$$g(t) = \frac{C}{1 + exp(-k(t-m))},$$

With C the carrying capacity, k the growth rate, m an offset parameter, g(t) the number of cases at any given time t. **[10]**

However, there are some limitations for only fitting a simple logistic growth model to the data. Sometimes the model only fits some stages of the covid-19 outbreak and cannot fit the overall trend well with only one function, especially when there are changepoints in the trend. The growth pattern of the cumulative confirmed or deceased cases may deviate from the logistic growth path due to some government policies. In order to find the changepoints in the spread of covid-19, we implement the logistic growth model with Facebook Prophet library. Prophet can model and forecast times series based on an additive model with three main model components: trend, seasonality and holidays, and there two build-in trend models: a saturating logistic growth model and a piecewise linear model.

We implemented Prophet's logistic growth model on the cumulative deceased cases for South Korea and the five states in the US with only weekly seasonality. We then limited the model to have up to two changepoints to avoid overfitting and increased the changepoint prior scale to 0.5 to make Prophet be able to find the best changepoints in a larger time space. The capacity (cap) for each logistic growth model was estimated from fitting a logistic growth function to an early stage of the data and each cap was increased a little to make sure the model wouldn't reach the cap within a 30-day prediction window. The model was trained by employing Maximum a posteriori (MAP) estimation and the results can be seen in Figure 4.

Changepoints: South Korea

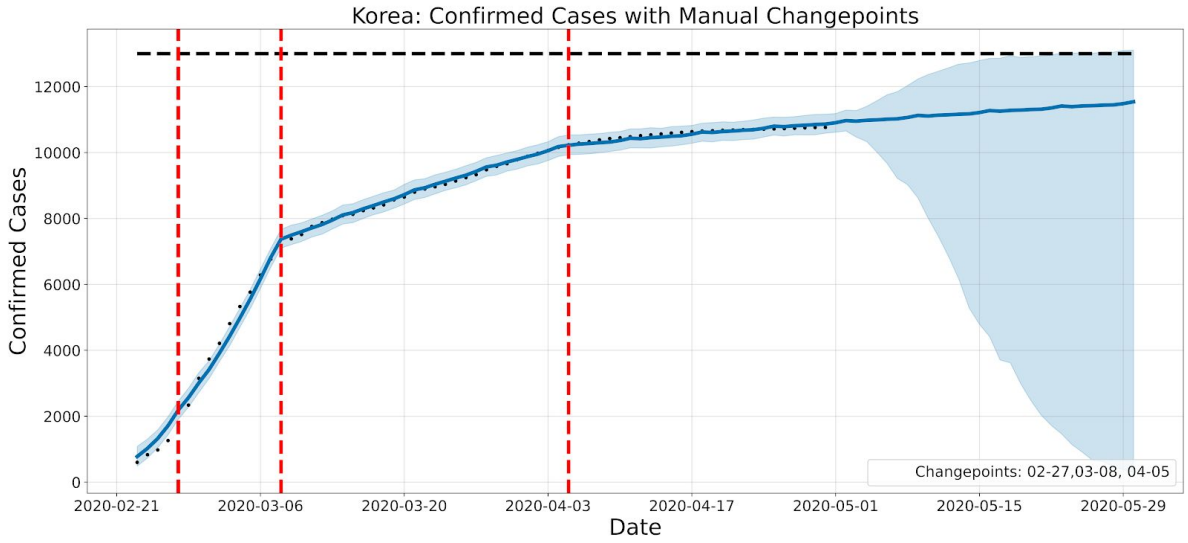

Figure 3: Facebook Prophet model of confirmed cases when changepoints ['2020-02-27','2020-03-08','2020-04-05'] were manually added. 4/05 corresponds to approximately 2 weeks after Daegu (the city primarily hit by Covid-19) started "enhanced social distancing" similar to U.S. shelter in place, and 2 weeks after the peak in new deaths. Voluntary self-quarantine was started in Daegu on 2/23, ~2 weeks before 3/08 declining changepoint. The 2/27 changepoint we believe is tied "patient 31", discussed below.

Patient 31

This group believes the changepoint on February 27th in SK was due to "patient 31", a member of the Shincheonji Church. This patient continued to go to church gatherings after being tested positive on February 21st, and cases quickly climbed. The church would at one point account for the majority of cases in the country. Officials quickly targeted the church for extra testing,

which fits in with the onset of cases seen at this time **[11].** This exposes the impact of random human nature and the importance of social distancing.

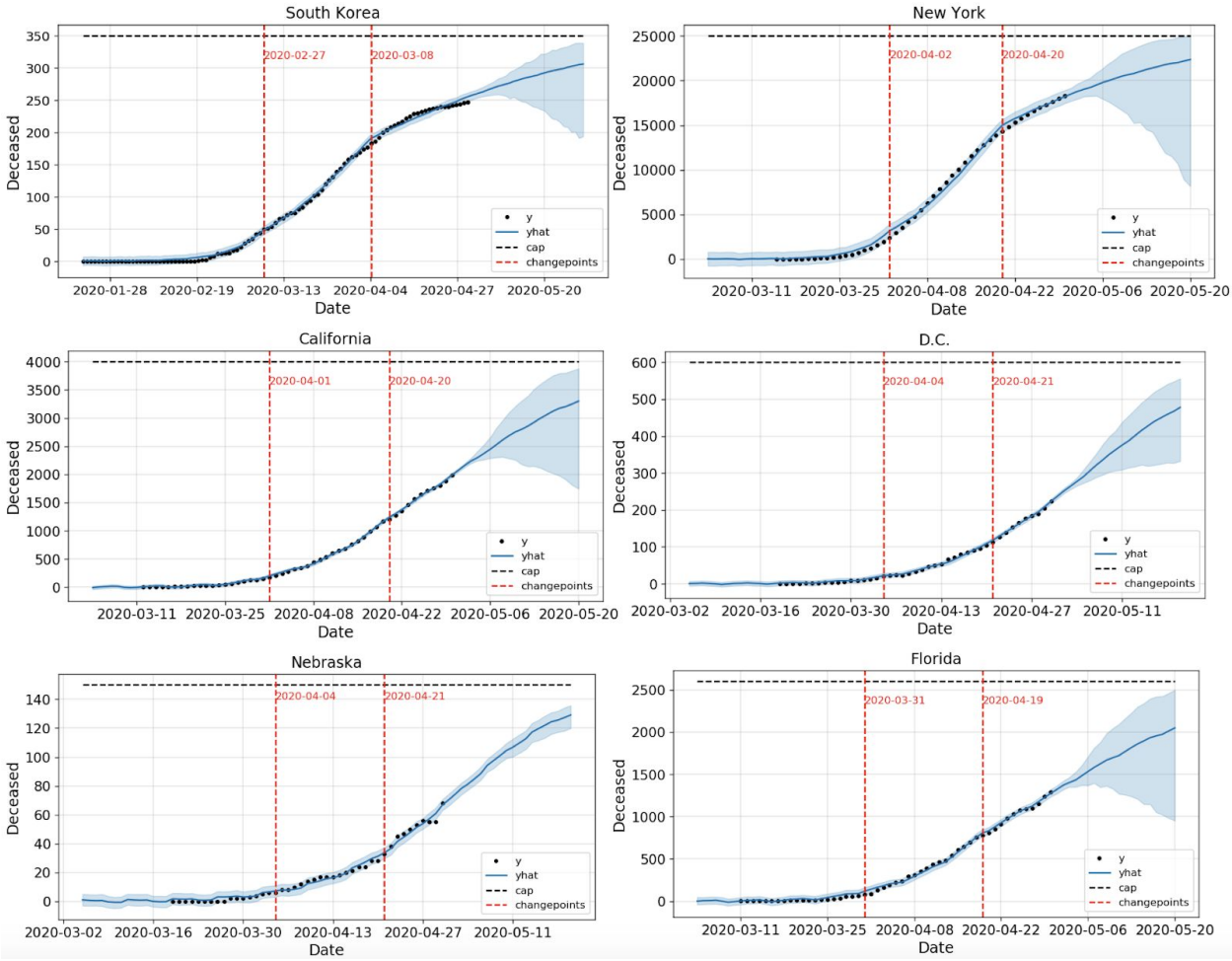

Figure 4: The Facebook Prophet fits with a logistic growth model applied to cumulative deceased cases in South Korea and five U.S. states: New York, California, Washington D.C., Nebraska and Florida. The changepoints for South Korea were manually added as in figure 1 and the changepoints for the five states were determined by employing MAP estimation in the Prophet fitting.

Additionally, we used common distance metrics such as the Minkowski metric to quantify similarity between deaths-over-time curves in five U.S. states and in South Korea. The Minkowski metric is defined as:

$$D(X,Y) = \left(\sum_{i=1}^{N} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Where p=1 corresponds to the *Manhattan* distance, and p=2 corresponds to the *Euclidean* distance. The Pearson Correlation Coefficient is useful for comparing trends between time series, but not very insightful in this case seeing as the cumulative deaths-over-time curves are all monotonically increasing functions.

Another method that we implement to give us a rudimentary measurement of synchrony is the time lagged cross-correlation. For discrete time series, f(t) and g(t) the function the operation is defined as

$$(f \star g)[n] \equiv \sum_{m=-\infty}^{\infty} \overline{f(t)}\, g(t + \tau).$$

Essentially the lag time, τ, repositions g(t), and at a given alignment the correlation between the time series is measured, and this is done with both a forward and backward lags. We measure the cross-correlation between South Korea and the 5 states, of cumulative tests and cumulative deaths. Prior to measurement, the time series are called to units of per capita via their respective population sizes and are standardized.

Conclusions

From the Prophet logistic growth model, we identified that New York state has a decreasing growth rate for the cumulative number of deaths since April 20th.The growth rate of the other four states are still increasing or almost reaching a constant, which represents that they are in the middle stage of the outbreak. We also found that the development stage of covid-19 be longer than what a simple logistic growth model predicts because the growth rate ~remains constant when it reaches the peak. It could be due to the limitation of the daily testing numbers. Changepoints were manually used to link policy and events in the SK case. For the U.S., a pattern of changepoints for the first month of May and the end of May was found for most states. This could not be as directly attributed, since most states are still in the growth stage of cases. One hypothesis is the early changepoint is related to stay at home procedures (which most states implemented ~2 weeks prior). The fact that this is a growth changepoint indicates measures were taken reactively rather than proactively, and were too late in most cases.

We used Minkowski metrics to examine similarity between the deaths-over-time of 5 US states vs. South Korea and found a wide discrepancy in metrics between states that did not appear geographical in origin. We considered if differences in population density could explain why some states were closer to the South Korean curve than others. As shown in Table 2, the distances of the NY and NE curves are well explained by their high and low (respectively) average population densities, relative to South Korea's. However, FL & CA, despite having mean population densities lower than South Korea's, have both experienced significantly higher rates of death overall, indicated by their high Minkowski metrics. We tentatively connect this discrepancy to policies taken by local governments, such as Florida's government's decision to reopen beaches early **[12]**. One weakness of this analysis is that we are comparing the mean population density of many cities in each state against that of a single city in South Korea, Daegu. This could be improved by limiting the population density sampling of states to major metropolitan areas only.

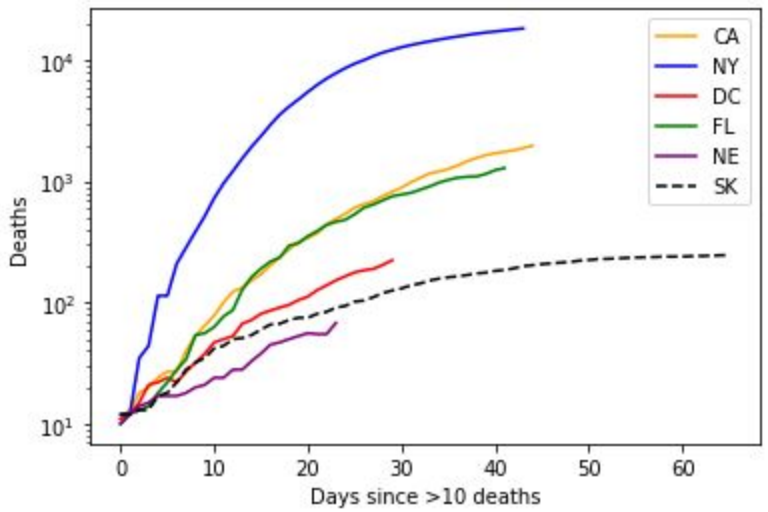

Figure 5: Number of deaths in 5 U.S. states vs. South Korea beginning from time of first 10 deaths.

	Minkowski (p=2)	**Minkowski (p=1)**	**Pearson R**	$\frac{\bar{\rho}_{state}}{\rho_{SK}}$
CA	5360.88	25273.0	0.978	0.76
NY	66059.58	332441.0	0.989	2.45
DC	233.42	883.0	0.986	1.55
NE	85.749	355.0	0.973	0.44
FL	3384.84	16014.0	0.991	0.45

Table 2: Minkowski distances and PCCs for each U.S. state curve vs South Korea curve shown in Figure 5. Rightmost column is the ratio average population density of each state to the population density of Daegu, SK. Average state population density is given as the mean population density of all cities with over 100,000 residents.

	Normalized Cross-Correlation: Deaths		**Normalized Cross-Correlation: Tests**	
State	**τ - of max corr. (days)**	**Max corr.**	**τ - of max corr. (days)**	**Max corr.**
CA	-6	0.74	-9	0.73
DC	-3	0.53	-16	0.63

FL	-4	0.71	-12	0.72
NE	-4	0.44	-18	0.54
NY	-4	0.74	-9	0.80

Table 3: The tables lists the lags of maximum correlation between the U.S. state time series and corresponding normalized-correlation value. This is done with deaths initiated after the 10th death and tests initiated after the 100th positive case.

From this measurement, we find the lag time, for which the state time series are most similar to the corresponding ones of South Korea. For example, if we look at testing in New York, we notice the max correlation is at a lag of 9 days and with a correlation of 0.8: essentially, when we shift the New York testing curve backward in time 9 days it would be highly similar to that of S. Korea. There is some similarity between NY at t = 9 days after the 100th case and S. K. at time 0. Overall, we can to some extent see that testing in these states have kept apace compared to earlier in the pandemic.

The same visual analyses can be done with the deaths as well, however a clear overlaying of the U.S.death-curves with South Korea proves more infeasible. However, this introduces the limitations of this basic correlation approach, for each measurement of correlation is global, and thus lacks dynamical details. So for deaths, the max correlation is most likely going to depend on the high similarity of accelerated portions of the curves while minimizing portions that anti-correlate. The loss in correlation would come from S.K.'s deaths slowing down, while the U.S still has a significant interval of rapid increase.

One other interesting thing to note about the time-lagged, discussed in the next section, cross correlation, is that when implemented between the new daily positive cases and new daily deaths, its able to compute the death lag close to the Korea's, with the τ = for maximum correlation being 23 days, but only correlation of 0.58. It however does not return the same accuracy of the U.S., for example Nebraska max was at τ = 3 days and τ =- 4 days for D.C.. An explanation as to why we see South Korea's death lag so accurately could be because of their combined testing, isolation and or hospitalization of infected individuals, and the isolation of others via contact tracing. The situation in the U.S. is more variable and less thorough. Then there is the noise given that not all infected individuals will die--it significantly varies by condition--or die in the same interval, so this may not work well with a process that relies on globality.

Another interesting parameter is something we called "death lag". There is a strong resemblance in new confirmed cases and deaths, if you shift the death curve into the future and scale it up. A larger shift means in general there is a longer time between patients being confirmed positive and dying. A larger scale factor means less of those tested positive die. For SK the shift is 24 and the scale is 80 (so ~1 in 80 tested die). Of note is California's low shift, likely due to the vulnerable elderly population being hit first. Also New York's scale of 14 fits early, but falls off, indicating death rates declined with time.

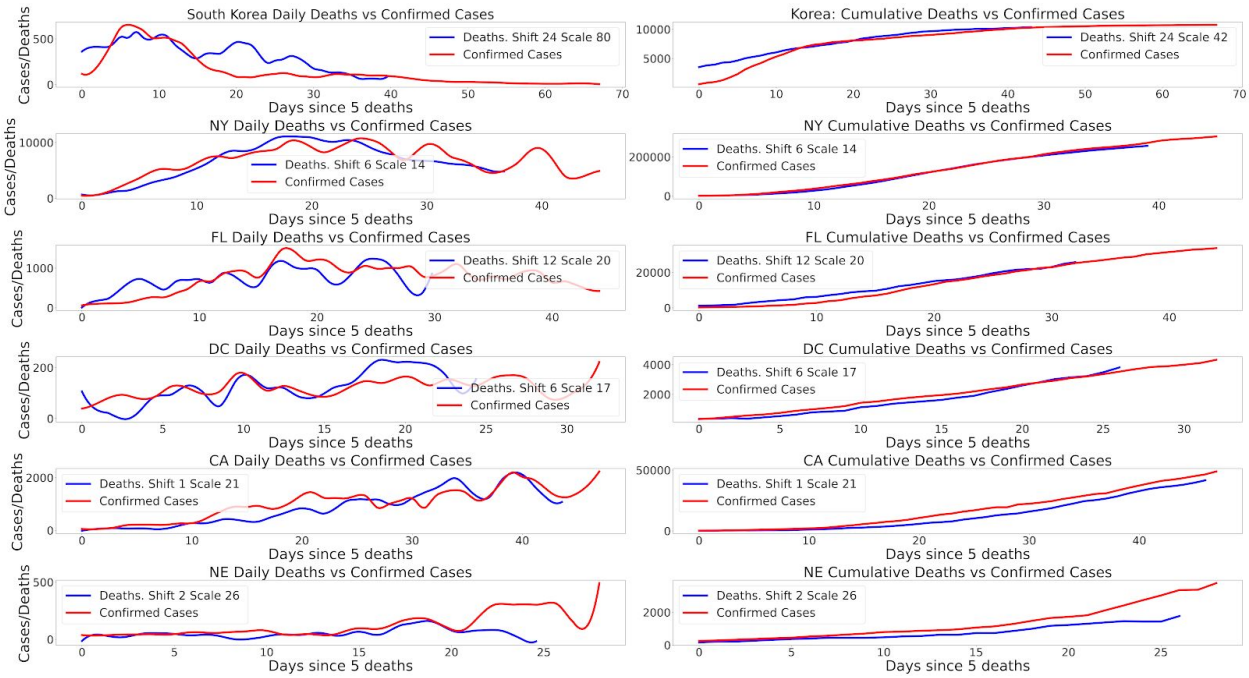

Figure 6: Death lags for South Korea and states. Left column is comparing new cases vs new deaths, the right column is cumulative. The results shown are from running interpolation and then passing the data through a 3rd order Savitzky–Golay filter. South Korea required a shift of 24 days into the future for new deaths to overlap new cases, with a scale of 80, by far the largest out of all regions. California had the smallest shift (indicating quick death after confirmation), while New York had the smallest scale factor (indicating a larger proportion of deaths).

We inspect the aspects, primarily death, of the Covid19 pandemic in South Korea and 5 U.S. locations with change point analytics, inspection of lag effects, and synchrony measurements. Overall we see that for the US to have the same outcome as South Korea, fast and aggressive testing and isolation is essential. But this project is rudimentary and has components that can be improved with newer data. The literature supports temperature and humidity as viable covariates to the spread of the virus, based on prior pandemics, and such data could be incorporated into future analyses. Even within our original ensemble of locations, we have climate diversity. If we choose to move forward with synchrony measurements such as time- lag cross correlation, we will have to invest in methodologies that go into a more detailed and dynamical assessment of the correlation.

Github Link

https://github.com/RileyWClarke/MLTSA_COVID19

Bibliography

[1] Division of Risk assessment and International cooperation, KCDC; *The updates on COVID-19 in Korea as of 5 April* (2020). Retrieved From: **(https://www.cdc.go.kr/board/board.es?mid=a30402000000&bid=0030)**

[2] Fisher, M., Sang-Hun, Choe; *How South Korea Flattened the Curve* (2020). The New York Times, Section A, page 12. Retrieved From: (https://www.nytimes.com/2020/03/23/world/asia/coronavirus-south-korea-flatten-curve.html).

[3] Roser, M., Richie, H., Ortiz-Ospina, E.; *Coronavirus Disease (COVID-19) – Statistics and Research* (2020). Published online at OurWorldInData.org. Retrieved From: https://ourworldindata.org/coronavirus

[4] Almukhtar et al; *District of Columbia Coronavirus Case Count* (2020). The New York Times. Retrieved From: https://www.nytimes.com/interactive/2020/us/washington-dc-coronavirus-cases.html

[5] Katz, J, Lu, D., Sanger-Katz, M; *U.S. Coronavirus Death Toll Is Far Higher Than Reported, C.D.C. Data Suggests (*2020). The New York Times. Retrieved From: https://www.nytimes.com/interactive/2020/04/28/us/coronavirus-death-toll-total.html

[6] UC San Diego Health; *Coronavirus Information* (2020). UCSC School of Medicine. Retrieved From: https://health.ucsd.edu/coronavirus/Pages/FAQ.aspx

[7] Chowell, Gerardo. "Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts." *Infectious Disease Modelling* 2.3 (2017): 379-398.

[8] Wu, Ke, et al. *Generalized logistic growth modeling of the COVID-19 outbreak in 29 provinces in China and in the rest of the world* (2020). Published online at *arXiv preprint arXiv:2003.05681*. Retrieved From: https://arxiv.org/abs/2003.05681

[9] Taylor SJ, Letham B.; *Forecasting at scale* (2017). Published online at PeerJ Preprints 5:e3190v2. Retrieved From: https://peerj.com/preprints/3190/

[10] Wu, Ke, et al. *Generalized logistic growth modeling of the COVID-19 outbreak in 29 provinces in China and in the rest of the world* (2020). Published online at *arXiv preprint arXiv:2003.05681*. Retrieved From: https://arxiv.org/abs/2003.05681

[11] Introvigne,M., Fautré, W., et al.; *Shincheonji and Coronavirus in South Korea:*

Sorting Fact from Fiction (2020). Published online at cesnur.org. Retrieved From: https://www.cesnur.org/2020/shincheonji-and-covid.htm

[12] Weisfeldt, S., Flores, R.; *More Florida beaches are reopening* (2020). CNN. Retrieved From: https://www.cnn.com/2020/04/24/us/florida-beaches-reopening-coronavirus/index.html