ISYE 6739 Project Final Report

Lechen Yu, Sara Miller, Matt Gudorf, Luke Erlandson

Spring 2020

Abstract

This work investigates the geographic and temporal transmission of the 2019 novel coronavirus. A variety of statistics tools were used to draw inferences on this highly-dynamic problem. Social distancing, quarantines, stay-at-home orders and other mandates have slowed the spread of COVID-19 throughout the world. These practices come with their own costs, however, in regards to mental health, unemployment, education, etc. Because the situation is in a constant state of change, it is important to identify the most effective measures because they might need to be reenacted upon the emergence of a second wave of proliferation. We preform an investigation on the doubling rate of the infection rate by country, both as a distribution and looking at individual countries. Finally, a literature review of peer-reviewed medical journals was performed to compile a data set with published estimates of the basic reproduction number (R_0). From these published estimates, a meta-analysis was performed to answer the following question: is R_0 for COVID-19 in China decreasing with time? A simple linear regression analysis showed no significant correlation between the dates of the studies and their resulting estimates for R_0. Therefore, we conclude that more time must pass before the relationship between R_0 and time can be gleaned from simple linear regression. An online version of our presentation is available via Google Slides.

Contents

1 Background and Justification

The 2019 novel coronavirus (COVID-19) has already resulted in more deaths than severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) combined, so understanding the transmission dynamics of this outbreak is critical to informing the public health approach [1]. The current report focuses on the growth patterns associated with COVID-19, as indicated by the number of confirmed cases and number of deaths. To understand this phenomenon, we consider the effect of weather and geographic location, estimates of the virus' reproduction number obtained from medical journals, the doubling time between countries, and the impact of self-isolation measures.

2 Description of Raw Data

The COVID-19 Open Research dataset [33] is used in the study of COVID-19's spread. This dataset consists of there parts: a) COVID-19 confirmed cases and fatalities group by regions (states), b) geography data (e.g., latitude and longitude), and c) 9 categories of weather data imported from NOAA GSOD dataset [34]. It records the confirmed cases from 2020-01-22 through 2020-04-11. To reason about the spread of COVID-19 in China, we analyzed the average growth rate and daily growth rate in each region, via a number of statistical methods. The average growth rate and series of daily growth rates are calculated based on the number of confirmed cases. The normality of daily growth rates is tested by D'Agostino's K-squared test [35]. For each region's average growth rate, two analysis was conducted to figure out potential factors that have a positive/negative effect on the spread of COVID-19: a) using covariance to analyze the relationship between the growth rate and distance to Hubei (the outbreak province of COVID-19), and b) using ANOVA to compare the growth rates in distinct environments.

Three datasets were used in the test of the effects of country and government mandates. These three datasets included data on the dates of different quarantine measures per country [6], the time series data for the case numbers [4] and the time series data on the number of tests [5]. These datasets are quite inconsistent due to differences in reaction to the pandemic and the inconsistency of reporting. This manifests as irregular time-series, in terms of the dates on which they are defined, as well as missing values. Because of this, a number of actions which may or may not be heavy handed needed to be employed. These actions accounted for the following: differences in time series ranges, missing values, quarantine measures which occur before the first recorded case in a country, dissimilar quarantine measures taken and differences in the countries whose data was recorded. To make the data uniform, the time series were normalized to be from December 31st 2019 to April 20th, 2020. The end date was simply because one of the data sets has not been updated in the past week. The original missing values and those created by this normalization were handled by using linear interpolation on the interior of the time series and linear extrapolation for the beginning of the time series. This is only applied at the very beginning of each countries time series (if values are missing) typically when the cases number is very small or even equal to 1. Therefore, I believe this action is justified by using linear expansion of exponential growth. After the time series were made uniform, the specific government responses to investigate were chosen by how widespread their adoption was. The countries were determined by taking the intersection of all countries present in the three data sets.

The doubling time of the virus is an important number as it provides an indication of how quickly the spread is happening. It is relatively simple to determine from cumulative data, directly arising from the data rather than requiring additional analysis. However, it does not provide specific or individual information, and can vary greatly depending on the timescale. To calculate and analyze the doubling of COVID-19, we looked at a dataset provided by Johns Hopkins [3], which provides information of many different regions on a day by day basis. These regions are either global regions, consisting of countries or large regions. Additionally, data for the United States in the format was used. This data is updated daily, as such we have data from 2020-01-22 through 2020-04-24 The data consists of a CSV file where each row corresponds with a region, there are a few columns of meta-data, followed by the remaining columns, one for each day since 2020-01-22, which contain the number of cumulative confirmed cases on a given day. Additionally, we referenced the calculated doubling time from [7], [8], to compare our results to.

The basic reproduction number, R_0, is used by epidemiologists to quantify the intensity of an infectious disease. A disease's R_0 is defined as the average number of people who will contract a disease from one contagious person. Thus, an accurate estimate of R_0 is instrumental to characterizing the transmission of COVID-19. An $R_0 > 0$ indicates that the number of infected people is likely to increase, whereas an $R_0 < 0$ indicated that the transmission of the disease is likely to die out. Therefore, the higher a disease's R_0, the greater the risk to a population. A literature review of peer-reviewed medical journals was conducted to compile a data set of estimates for R_0 for COVID-19 in China. A total of 32 estimates for R_0, along with the corresponding 95% confidence intervals and study dates, are provided in Appendix A. In the literature, a confidence level of 95% was found to be the standard for discussions involving R_0. The "Study Date" column in Appendix A refers to the cutoff date for the number of COVID-19 confirmed cases that was used in each study. Some researchers used different mathematical models to develop multiple estimates of R_0 and each of those estimates was included as separate data points. Due to the overlap of data sets, researchers, and mathematical approaches to estimating R_0, this data is not independent.

3 Statistical Analyses

3.1 COVID-19 growth interaction with geography and weather data

We leveraged the confirmed cases in the COVID-19 Open Research dataset [33] to calculate the daily growth rate and average growth rate. The number of confirmed cases in each region is treated as a random variable, and the recorded value in the dataset is a set of random samples. For each region in China, we used Equation (1) to calculate the daily growth rates from 2020-01-22 through 2020-04-11, and Equation (2) the average growth rate over this period. We refer to the number of confirmed cases on the ith day after 2020-01-22 as $Case_i$.

$$GrowthRate_i = \frac{Case_i - Case_{i-1}}{Case_{i-1}} \tag{1}$$

$$AverageGrowthRate = (Case_n - Case_1)^{\frac{1}{n-1}} - 1 \tag{2}$$

Figure 1 presents the average growth rate. To test its normality, we ran a D'Agostino's K-squared test. The test statistic's result is 18.69098315319971, and the corresponding p-value is 8.735838117897205e-05. It indicates that the average growth rate in the collected samples does not satisfy normal distribution.

Next, we examined the correlation between the average growth rate and the geographical distance to Hubei, the outbreak city of COVID-19. Our conjecture is the average growth rate decreases with an increase in distance. The geographical distance is calculated via the Geopy python library [36]. Figure 2 shows the average growth rate by distance to Hubei. The trend matches our expectations. For regions within 1000 km from Hubei, their average growth rate is greater than those distant regions. However, we still observed some counterexamples, for instance, Heilongjiang. Heilongjiang is China's northernmost province which shares a border with Russia, while Hubei is a landlocked province in central China. To figure out possible explanations for Heilongjiang's high average growth rate, we did a literature research for COVID-19 news in Heilongjiang. We found that since early April, there were a large number of imported COID-19 cases from Russia [37]. It is highly possible that these imported cases result in an increase of COVID-19's growth rate.

Apart from Figure 2, we also calculated the covariance matrix for the average growth rate and distance (we cannot apply ANOVA here since the average growth rate is not normally distributed). The result is shown in Equation (3), where the negative covariance indicates a negative correlation between the two variables.

$$cov(AverageGrowthRate, Distance) = \begin{bmatrix} 1.58403777e-04 & -2.48454066e+00 \\ -2.48454066e+00 & 3.08315655e+05 \end{bmatrix} \tag{3}$$

The last analysis we conducted on the COVID-19 Open Research dataset is the ANOVA between the daily growth rates and weather data. We chose four regions for this analysis: Beijing, Guangdong, Liaoning,
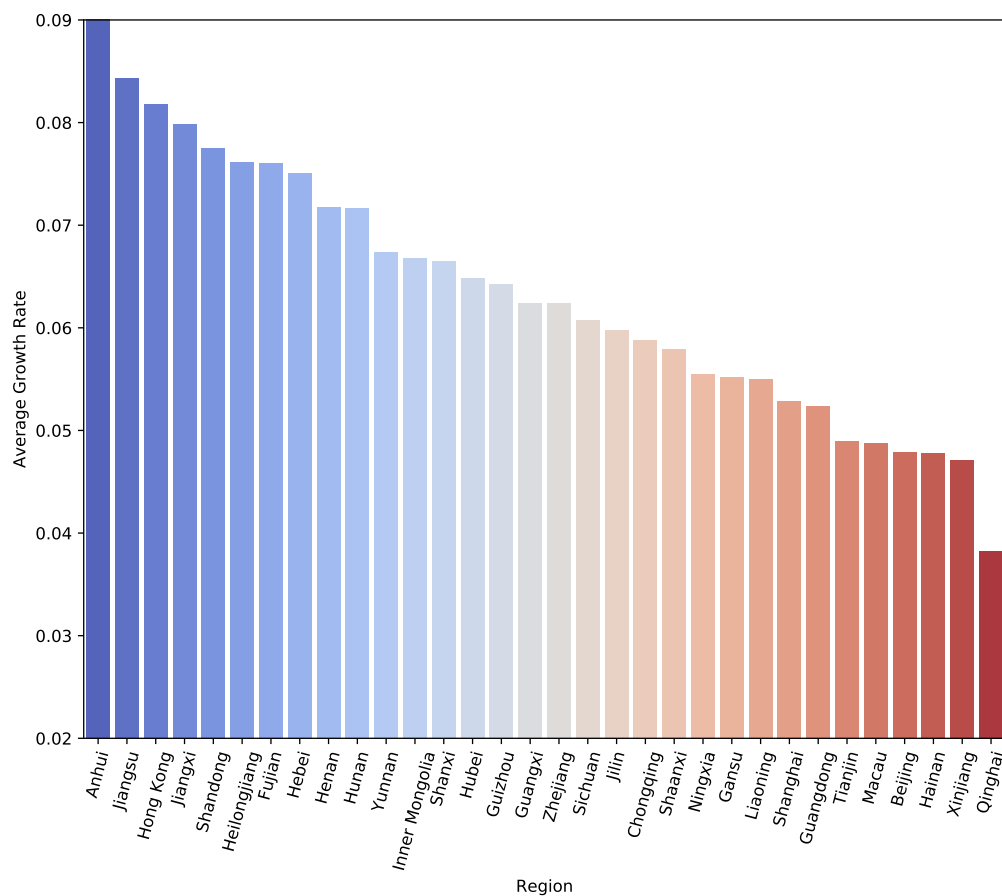
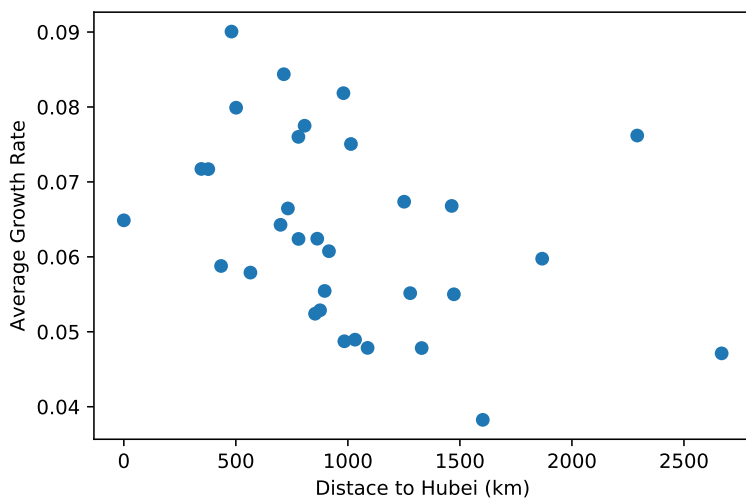Figure 1: Average Growth Rate of COVID-19 in China

Figure 2: Average Growth Rate by Distance to Hubei

4

Table 1: ANOVA Result for Mean Temperature and Total Precipitation

Measure	sum_sq	df	F	PR(>F)
C(temp)	8.453343e-17	2.0	7.406908e-16	1.000000
C(prcp)	3.274350e-18	2.0	2.869020e-17	1.000000
C(temp):C(prcp)	4.702349e-02	4.0	2.060124e-01	0.813969
Residual	1.352415e+01	237.0	NaN	NaN

and Shanghai. These four provinces are located in different areas in China, so that their weather data is distinct. The corresponding mean of the daily growth rate is labeled as μ_1, μ_2, μ_3, and μ_4. For each pair of weather data, a two-way ANOVA is carried out to verify the following hypothesis test:

$$H_0 : \mu_1 = \mu2 = \mu3 = \mu4 \tag{4}$$

$$H_1 : at\ least\ one\ mean\ differs\ from\ others \tag{5}$$

In Table 1, we display the result of ANOVA for the mean temperature and total precipitation. It demonstrates that there is no significant difference between growth with distinct mean temperature and total precipitation. For other weather data, the results are similar. Since Figure 2 also illustrates that the average growth rate is between 0.04 - 0.09, which is quite a small range, we can conclude that the weather data has little effect on the spread of COVID-19.

3.2 COVID-19 growth interaction with quarantine measures and country

With the world currently under maintenance and public sentiment deteriorating, it is important to identify the most efficient quarantine measures so that governments know how to be proactive going forward. Given a set of countries how do specific quarantine measures affect the growth rate of COVID-19? To analyze this, we used the strictly increasing quantity of total number of tests per capita. Th main external factor which was not taken into account is the fact that the more tests a country performs the more cases typically emerge; i.e., the rate over time may increase as a result of increased testing and not spread of the disease itself. Figure
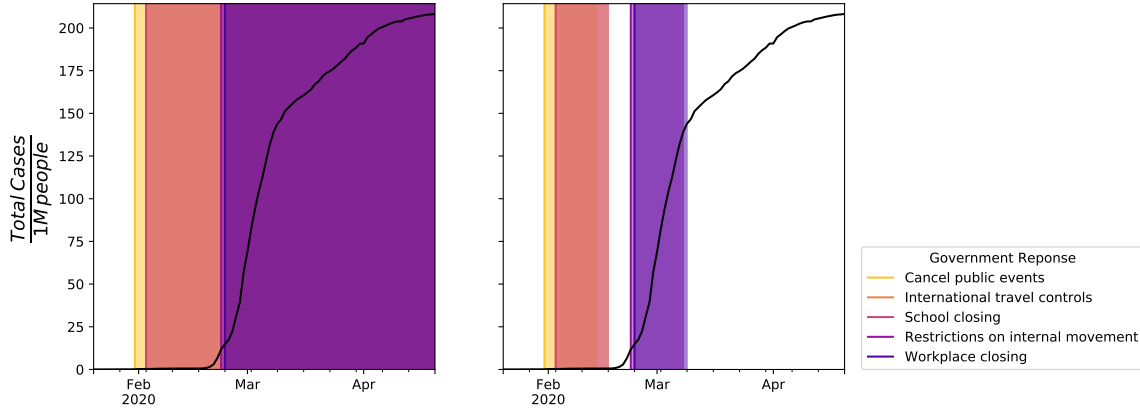
Figure 3: The time series of the total cases per capita for South Korea. (Left) The colored rectangles indicate the active dates for each quarantine measure. (Right) The width is now a 14 day window, approximating the incubation period for COVID-19.

3 demonstrates a time series of total cases for South Korea. The widths of the colored regions represent different quantities; on the left it demonstrates the total amount of time certain measures have been in effect while on the right it only shows a 14-day window after the implementation of each measure. These 14-day windows attempt to account for the time-delayed effect resulting from incubation time. The claim that we are pushing is that the rate of change increased until 14 days after the strictest level of quarantine measures were introduced, implying that only then were the measures sufficient. This behavior is incorporated directly

into the analysis. For every quarantine measure, 14 day windows of the time series are taken prior and after each implementation date. The average of the total cases per capita is taken on these windows which results in a triplet of past, present, and future cases per capita values. By assuming exponential growth, the growth rate before the quarantine measure and after as well as their difference are calculated using

$$\rho_i = \frac{1}{\Delta t_{mi}} \log(\frac{\phi_m}{\bar{\phi}_i})$$

$$\rho_f = \frac{1}{\Delta t_{fm}} \log(\frac{\bar{\phi}_f}{\phi_m})$$

$$\rho = \rho_f - \rho_i \,, \tag{6}$$

where over-bars denote averages over each window. By iterating over the countries and government responses, equation (6) produces a table of observations on which two-way ANOVA can be applied. Specifically, the blocks were taken to be the different quarantine measures and the treatments were taken to be the countries. The confidence level is chosen to be 95% and the corresponding hypothesis set can be written as

$$H_0 : \mu_0 = \mu_1 = ... = \mu_a$$
$$H_1 : \text{any of the means is different}$$

The actual application of ANOVA was straight forward using `statsmodels` Python API. The results in table 2 look promising with p-values of 0.037 and 0.000003 for the choice of quarantine measure and country, respectively.

Table 2: Two-Way ANOVA results for change in growth rate by factor

	Sums of squares	d.o.f.	F-stat	p-value
Type	6.757982	4	2.618791	0.037044
Country	71.847056	40	1.8343	0.000003
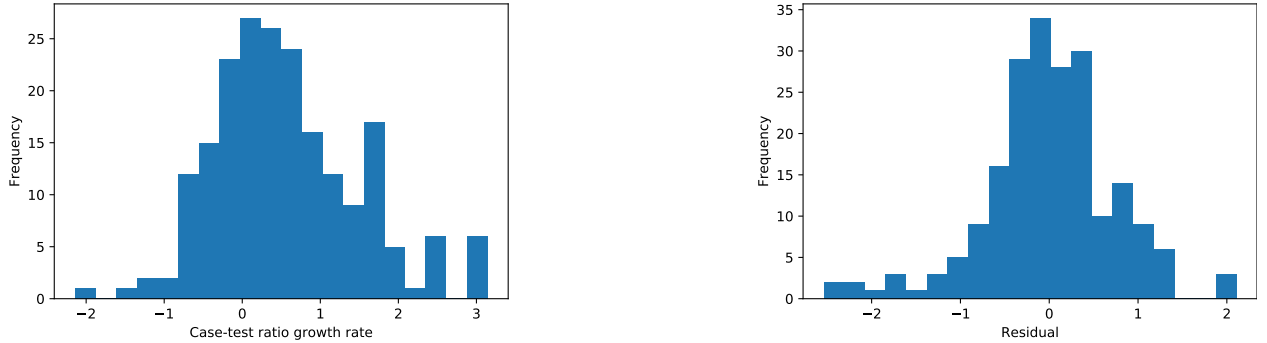
Figure 4: Observed value and residual distributions

To validate these results the distribution of the observations and their residuals (Figure 4) as well as variance comparison plots for countries and government responses are displayed (Figure 5). The residuals and observations are approximately normally distributed and the variances for the blocks seem similar. The main area of concern is the variances of each countries' residuals. These disparate variances affect the interpretation of the ANOVA results; especially the effect that the countries have on the rate change. Additionally, these results may be misleading because the quarantine measures overlap with one another; when combined with the time delayed nature of the problem these results should be viewed skeptically.

3.3 Doubling Time Analysis

To analyse the spread of Covid-19, we looked at the stasticial significance of the doubling time between different countries. To accomplish this, we used the dataset from John Hopkins, which included cumulative

Figure 5: Variance comparisons for treatment residuals (left, countries) and block residuals (right, government responses). There are too many countries to label conveniently but the order from left to right is equivalent to the lexicographical ordering in the corresponding appendix table.

Table 3: Doubling time of countries

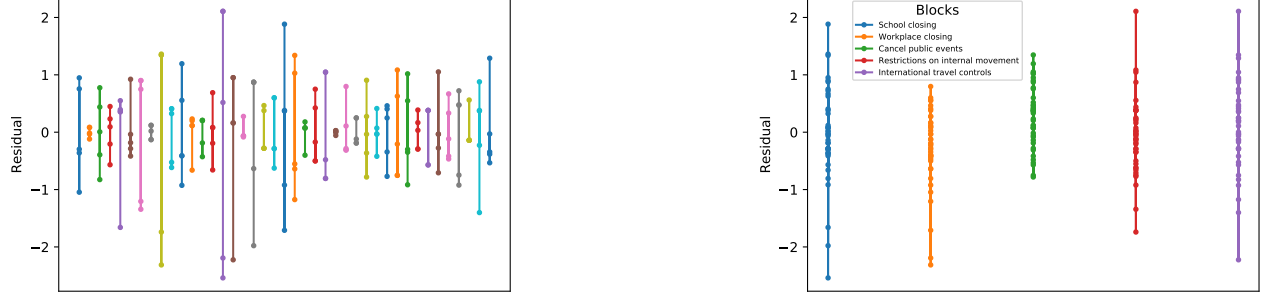| Country/Region	Days	Country/Region	Days	Country/Region	Days	Country/Region	Days
Bangladesh	3.54	Ireland	5.78	Sweden	6.64	Nigeria	8.27
Turkey	3.98	Saudi Arabia	5.79	Morocco	6.66	Cote d'Ivoire	8.45
US	4.05	United Arab Emirates	5.90	Moldova	6.74	Croatia	8.50
Russia	4.17	Italy	5.95	Philippines	6.86	Thailand	8.62
Belarus	4.37	Netherlands	6.03	South Africa	7.27	Finland	8.77
United Kingdom	4.89	Dominican Republic	6.04	Oman	7.29	Qatar	9.00
Spain	4.91	Israel	6.11	Argentina	7.37	Kuwait	9.06
Brazil	5.02	Pakistan	6.13	Czechia	7.41	Denmark	9.06
Peru	5.08	Poland	6.14	Hungary	7.60	Japan	9.16
Canada	5.09	Chile	6.24	Algeria	7.67	Ghana	9.27
India	5.24	Iran	6.24	Afghanistan	7.91	North Macedonia	9.43
France	5.27	Romania	6.45	Azerbaijan	7.91	Armenia	9.66
Germany	5.29	Panama	6.47	Egypt	7.95	Slovakia	10.12
Ukraine	5.33	Colombia	6.57	Cameroon	7.97	Iraq	10.36
Portugal	5.50	Indonesia	6.61	Cuba	8.00	Bahrain	10.39
Ecuador	5.56	Uzbekistan	6.62	Singapore	8.12	Bulgaria	11.25
Belgium	5.69	Kazakhstan	6.64	Malaysia	8.24	Greece	11.36

Table 4: Doubling time of states

Province_State	Days	Province_State	Days	Province_State	Days	Province_State	Days
New York	3.54	Louisiana	4.57	South Dakota	5.18	Utah	5.79
New Jersey	3.68	Ohio	4.61	North Carolina	5.18	District of Columbia	5.96
Pennsylvania	3.90	Georgia	4.72	Tennessee	5.29	Kansas	6.29
Michigan	3.91	Arizona	4.84	Delaware	5.34	Washington	6.32
Illinois	4.16	Rhode Island	4.94	South Carolina	5.47	New Mexico	6.35
Massachusetts	4.26	Virginia	4.95	Wisconsin	5.50	Idaho	6.41
Connecticut	4.28	Missouri	5.01	Nevada	5.55	Arkansas	6.61
Indiana	4.34	Mississippi	5.08	Iowa	5.62	New Hampshire	6.79
Texas	4.46	Alabama	5.12	Oklahoma	5.64	Minnesota	6.80
Maryland	4.47	Colorado	5.16	Kentucky	5.74	West Virginia	7.00
Florida	4.48	California	5.17	Nebraska	5.78	Oregon	7.23

cases for different cities in different regions. We accumulated the data by region, before remapping the data to consider the number of days since the country first had 100 confirmed cases. This allows a better inter country comparison. We then filtered out countries for which had less than 1000 cases, to ensure that enough of the trend could be seen to derive some conclusions from, as well as countries who's reported cases did not increase more than 1% between the five most recent days (removing Australia, China, Iceland, Serbia,

(a) Linear Scale   (b) Log Scale

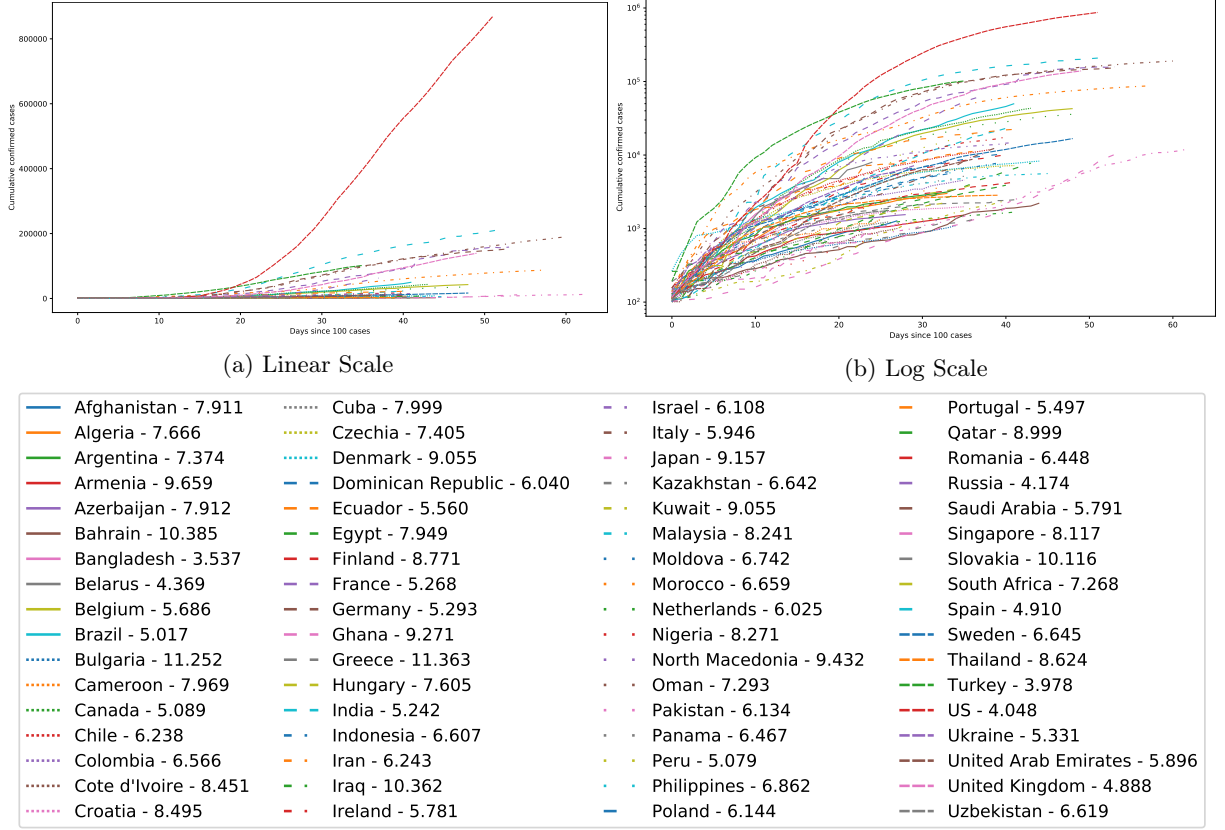|			
Afghanistan - 7.911	Cuba - 7.999	Israel - 6.108	Portugal - 5.497
Algeria - 7.666	Czechia - 7.405	Italy - 5.946	Qatar - 8.999
Argentina - 7.374	Denmark - 9.055	Japan - 9.157	Romania - 6.448
Armenia - 9.659	Dominican Republic - 6.040	Kazakhstan - 6.642	Russia - 4.174
Azerbaijan - 7.912	Ecuador - 5.560	Kuwait - 9.055	Saudi Arabia - 5.791
Bahrain - 10.385	Egypt - 7.949	Malaysia - 8.241	Singapore - 8.117
Bangladesh - 3.537	Finland - 8.771	Moldova - 6.742	Slovakia - 10.116
Belarus - 4.369	France - 5.268	Morocco - 6.659	South Africa - 7.268
Belgium - 5.686	Germany - 5.293	Netherlands - 6.025	Spain - 4.910
Brazil - 5.017	Ghana - 9.271	Nigeria - 8.271	Sweden - 6.645
Bulgaria - 11.252	Greece - 11.363	North Macedonia - 9.432	Thailand - 8.624
Cameroon - 7.969	Hungary - 7.605	Oman - 7.293	Turkey - 3.978
Canada - 5.089	India - 5.242	Pakistan - 6.134	US - 4.048
Chile - 6.238	Indonesia - 6.607	Panama - 6.467	Ukraine - 5.331
Colombia - 6.566	Iran - 6.243	Peru - 5.079	United Arab Emirates - 5.896
Cote d'Ivoire - 8.451	Iraq - 10.362	Philippines - 6.862	United Kingdom - 4.888
Croatia - 8.495	Ireland - 5.781	Poland - 6.144	Uzbekistan - 6.619

Figure 6: Doubling and cumulative cases (Countries)

South Korea due to growth).

This resulted in 68 countries, of which the minimum time frame was 93 . To calculate the doubling time for each country, we compared the number of days between the first day of 100 cases and the current day, and the number of doubling between the two dates. The resulting countries, cumulative cases, and doublings can be seen in Figure 6.

Treating each country as a separate sample, we performed a confidence interval on the expected doubling time, as well as plotting a histogram and box plot, and normality analysis. We determined that the sample mean was 7.01 , with a confidence interval (6.58, 7.45) with $\alpha = 0.05$. This is comparable to the results seen in [7], where they reference papers with confidence intervals 6.4 to 7.4, and 2.9 to 4.6, where our results fall very close to the first, and larger than [8] where they found a confidence interval for the initial Wuhan outbreak to be 2.3-3.3. We do note that our data uses much more recent data (up until 2020-04-24), and considers all data reported, where the papers only consider data up to their publication. Through a normality consideration, we see that we get a p-value of 0.33 , indicating that we can reject the hypothesis that the data is normal with $\alpha = 0.05$.

We also looked at the distribution of doubling within the United States on a per state basis. The same initial filtering was done, which removed Alaska, American Samoa, Diamond Princess, Grand Princess, Guam, Hawaii, Maine, Montana, North Dakota, Northern Mariana Islands, Vermont, Virgin Islands, Wyoming due to not having more than 1000 cases. The new plot can be seen in Figure 9, where a normality p-value of 0.33 , indicating that the data does in fact follow a normal distribution.
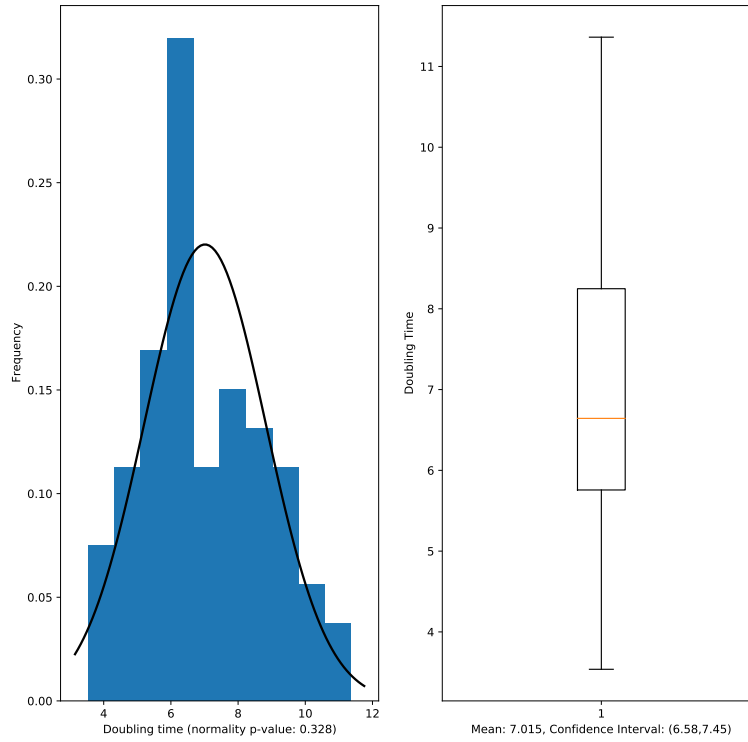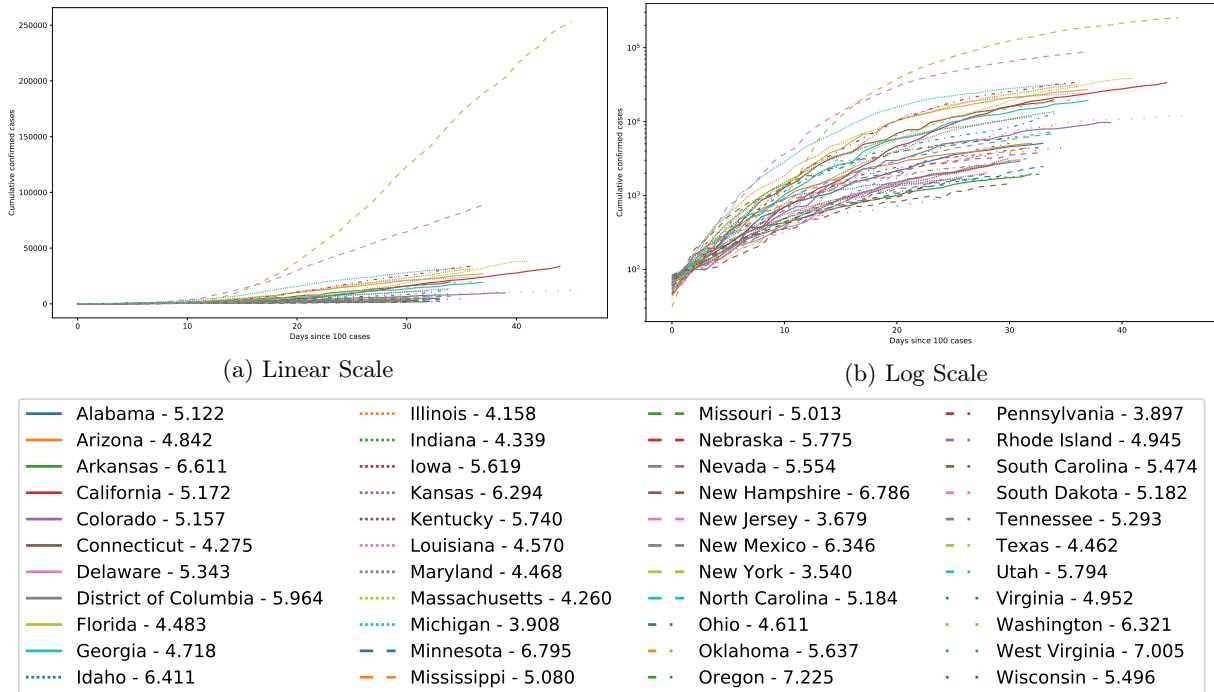
Figure 7: Histogram and box plot (Countries)

(a) Linear Scale

(b) Log Scale

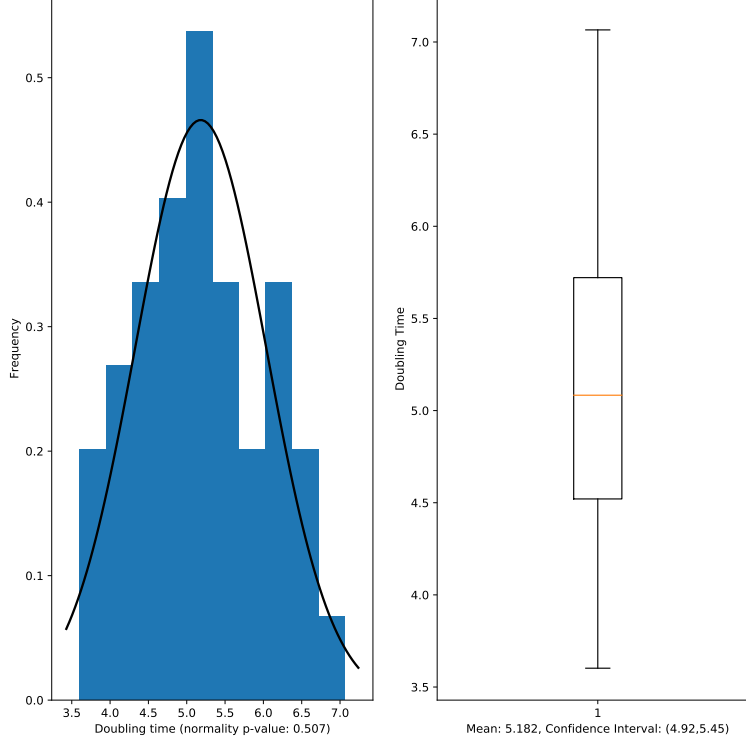|			
Alabama - 5.122	Illinois - 4.158	Missouri - 5.013	Pennsylvania - 3.897
Arizona - 4.842	Indiana - 4.339	Nebraska - 5.775	Rhode Island - 4.945
Arkansas - 6.611	Iowa - 5.619	Nevada - 5.554	South Carolina - 5.474
California - 5.172	Kansas - 6.294	New Hampshire - 6.786	South Dakota - 5.182
Colorado - 5.157	Kentucky - 5.740	New Jersey - 3.679	Tennessee - 5.293
Connecticut - 4.275	Louisiana - 4.570	New Mexico - 6.346	Texas - 4.462
Delaware - 5.343	Maryland - 4.468	New York - 3.540	Utah - 5.794
District of Columbia - 5.964	Massachusetts - 4.260	North Carolina - 5.184	Virginia - 4.952
Florida - 4.483	Michigan - 3.908	Ohio - 4.611	Washington - 6.321
Georgia - 4.718	Minnesota - 6.795	Oklahoma - 5.637	West Virginia - 7.005
Idaho - 6.411	Mississippi - 5.080	Oregon - 7.225	Wisconsin - 5.496

Figure 8: Doubling and cumulative cases (States)

Figure 9: Histogram and box plot (States)

3.4 Basic Reproduction Number Analysis

A meta-analysis was performed using published estimates for COVID-19's R_0. From the sample captured by the R_0 data set provided in Appendix A, descriptive statistics were used to quantify and visualize the central tendency and spread of these estimates. Table 5 summarizes some of the relevant numeric properties of the sample. Note that a mean R_0 of 3.45 is interpreted as an infected person transmitting the virus to 3.45 others, on average.
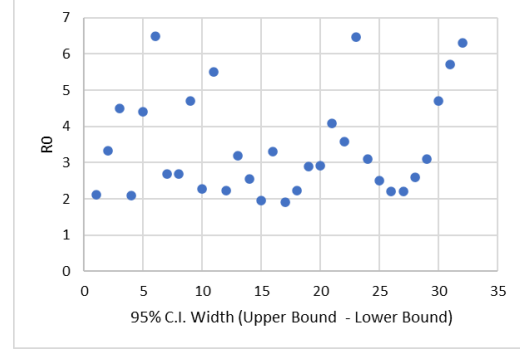
Table 5: Descriptive Statistics for the R_0 Estimates.

Measure	Result
Mean	3.45
Median (Q2)	3.01
Minimum	1.90
Maximum	6.49
Range	4.59
Standard Deviation	1.41
Variance	1.99
1st Quartile (Q1)	2.27
3rd Quartile (Q3)	4.43
Interquartile Range (IQR)	2.16

Using the quartile results in Table 5, a box plot of the R_0 estimates was plotted in figure 10 (a). To check for outliers, the lower bound and upper bound were calculated to be -0.96 and 7.66, respectively. Because all of the R_0 estimates used in the current analysis are contained within the range between the lower and upper bounds, none of the estimates in the data set were found to be outliers. To further visualize the spread of data, a scatter plot is provided in figure 10 (b). The magnitude of R_0 is plotted against the
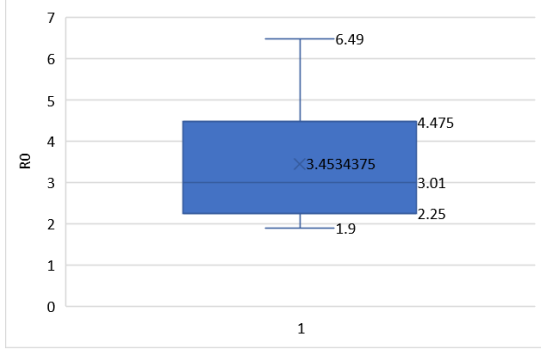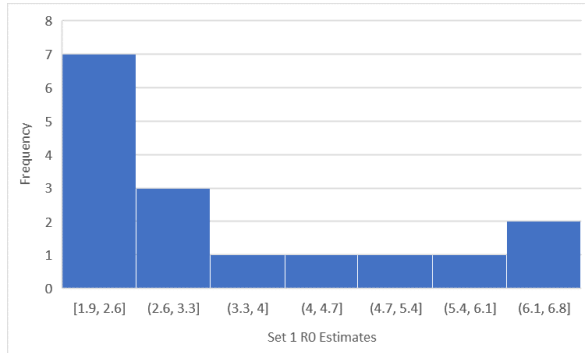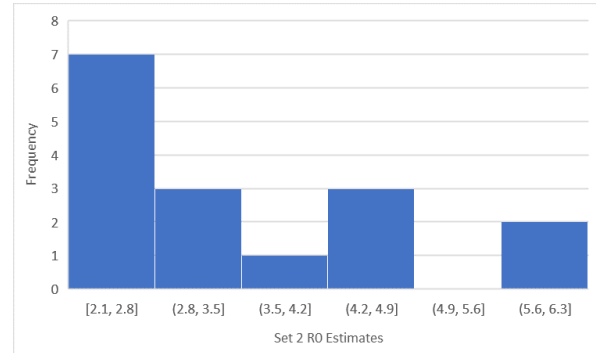
(a) R_0 Box Plot

(b) R_0 Scatter Plot

Figure 10: Descriptive Statistics for R_0 Estimates.

(a) R_0 Data Set 1 Histogram

(b) R_0 Data Set 2 Histogram

Figure 11: Distributions of Set 1 and Set 2 Data for R_0 Estimates

width of each estimate's 95% confidence interval, which was calculated by taking the difference between the confidence interval's upper and lower bounds.

Next, we aim to investigate the relationship between the magnitude of the R_0 estimate and the study date. The goal of this analysis is to indicate whether the transmission of the virus has grown or lessened over time. A Chinese governmental ban on public transportation that was invoked on January 23, 2020 for five major cities (Wuhan, Huanggang, Ezhou, Chibi and Zhijiang) [2]. Thus, we choose to divide the data set into two groups: set 1 includes R_0 estimates from studies that used COVID-19 case data leading up to and including January 23, 2020 and set 2 data includes R_0 estimates from studies that considered reported cases of COVID-19 since January 23, 2020. The sample size of set 1 and set 2 are equal with 16 estimates in each category. Histograms to show the distribution of both data sets are provided in figure 11. Both distributions are positively skewed, favoring lower magnitude R_0 estimates.

A simple linear regression analysis is performed using the study date as the independent/predictor variable and the R_0 estimate as the dependent/response variable. To turn the study date into an integer, the earliest study date was given a value of zero and each day thereafter is incremented by a value of one. The resulting regression equation is $\hat{y} = 0.00158X + 3.4152$. Because the slope of the regression line is nearly zero, we conclude that there is very little correlation between the study date and R0 estimate. The linear regression results are plotted in figure 12. A summary of the regression results is provided in table 6.
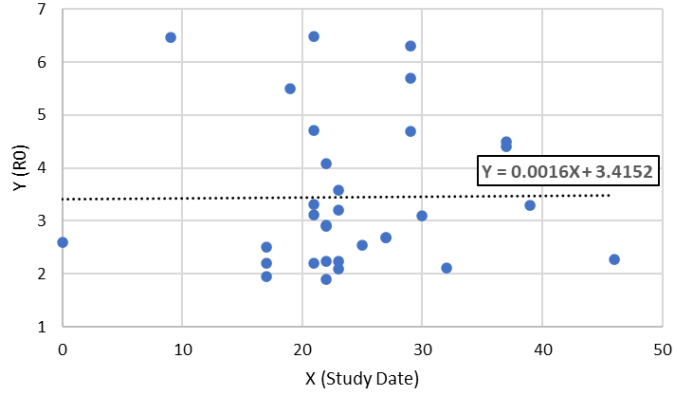
Figure 12: Linear Regression Results for R_0 Estimates

Table 6: Simple Linear Regression Results for R_0 Estimates.

Regression Result	Value
$\bar{X}$	24.16
$\bar{Y}$	3.45
$\hat{\beta}_0$	3.415
$\hat{\beta}_1$	0.0016
SS_{Error}	61.5468
SS_{Total}	61.5525
$SS_{Regression}$	0.0057
Means Squared Error (MSE)	2.0516
α	0.05
df (n - 2)	30
$t_{\alpha/2,n-2}$	2.042
Margin of Error of C.I. (E)	0.5337
95% C.I. for ŷ = 3.45	(2.9163, 3.9837)

4   Implications

Looking at the distribution of doubling time for various countries, a number of implications arise. First, a number of countries considered to be first world have short doubling time - notable US, the UK, Spain, Canada, France and Germany. This could be related to the amount of testing done, as well as how quickly the country responded as a whole. Secondly, as the doubling time does appear to follow a normal distribution, more advance analysis techniques could be used which require a normal distribution.

The initial hypothesis that R_0 would decrease over time in response to social distancing and restrictions on public gatherings was not confirmed in the current analysis. Instead, no significant correlation between study dates and estimated R_0 was found. Nonetheless, we do not conclude that these variables are entirely unrelated. Instead, it is proposed that more time and more observations are necessary to observe a relationship between these variables using simple linear regression. The implications of the two-way ANOVA analysis with countries and government quarantine responses are that both the factors have statistically significant effects. If the results are taken seriously then this would imply that a custom tailored approach should be considered when decided upon the quarantine measures to enact.

5 Future Work

The analysis on the effect of geography data indicates that it has a negative correlation with COVID-19's spread. We plan to further analyze this relationship in a fine-grained manner (e.g., analyzing the growth rate for each city in China). In addition, we will also examine other weather data to find potential factors that affect the spread of COVID-19.

The analysis on the effect of different quarantine measures do not take a number of details into account. The most important of which is the fact that the quarantine measures are not independent as they are overlap in regards to time. Therefore, to analyze quarantine measures correctly there should likely be a different method developed which accounts for this. Other considerations also need to be made such as how the number of cases depends on the number of tests.

Given the doubling rate per country, it would be possible to attempt to correlate factors of countries, such as population density, to see what factors impact the doubling time. Furthermore, one could look into how the doubling time changes with respect to the measures put in place as a proxy for determining the effectiveness of the measures.

Given more time, extending the literature review for published estimates of R_0 for COVID-19 in China to produce a greater sample size would strengthen the statistical analyses performed for the current meta-analysis. If enough observations were collected in each data set such that n $>$ 30 for estimates using reported cases prior to (data set 1) and after (data set 2) January 23, 2020, then using the assumption of approximately normally-distributed data could expand the number of inferences that could be drawn on this data.

References

[1] World Health Organization. Coronavirus disease 2019 (COVID-19) situation reports, 2020.

[2] BBC NEWS. "Coronavirus: Wuhan Shuts Public Transport over Outbreak." Jan. 23, 2020.

[3] Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)

[4] Our World in Data. Joe Hasell, Esteban Ortiz-Ospina, Edouard Mathieu, Hannah Ritchie, Diana Beltekian and Max Roser https://github.com/owid/covid-19-data/tree/master/public/data/

[5] Foundation for Innovative New Diagnostics. COVID-19 Testing Tracker. https://finddx.shinyapps.io/FIND_Cov_19_Tracker/

[6] A. Petherick, T. Hale, T. Phillips, S. Webster, Variation in Government Responses to COVID-19. Preprint https://www.bsg.ox.ac.uk/research/publications/variation-government-responses-covid-19

[7] M. Park, A. R. Cook, J. T. Lim, Y. Sun, and B. L. Dickens, "A Systematic Review of COVID-19 Epidemiology Based on Current Evidence," Journal of Clinical Medicine, vol. 9, no. 4, p. 967, Apr. 2020, doi: 10.3390/jcm9040967.

[8] S. Sanche, Y. T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, and R. Ke, "Early Release - High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2 - Volume 26, Number 7—July 2020 - Emerging Infectious Diseases journal - CDC," doi: 10.3201/eid2607.200282.

[9] Lai A, Bergna A, Acciarri C, Galli M, Zehender G. Early Phylogenetic Estimate of the Effective Reproduction Number Of Sars-CoV-2. Journal of Medical Virology. 2020.

[10] Tang B, Wang X, Li Q, Bragazzi NL, Tang S, Xiao Y, et al. Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. Journal of Clinical Medicine. 2020;9(2):462.

[11] Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of Travel Medicine. 2020.

[12] Imai N, Cori A, Dorigatti I, Baguelin M, Donnelly CA, Riley S, et al. Report 3: transmissibility of 2019-nCov. Reference Source. 2020.

[13] Riou J, Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019-nCoV. bioRxiv. 2020.

[14] Zhou C. Evaluating new evidence in the early dynamics of the novel coronavirus COVID-19 outbreak in Wuhan, China with real time domestic traffic and potential asymptomatic transmissions. medRxiv. 2020.

[15] Shao N, Cheng J, Chen W. The reproductive number R0 of COVID-19 based on estimate of a statistical time delay dynamical system. medRxiv. 2020.

[16] Shen M, Peng Z, Xiao Y, Zhang L. Modelling the epidemic trend of the 2019 novel coronavirus outbreak in China. bioRxiv. 2020.

[17] Read JM, Bridgen JR, Cummings DA, Ho A, Jewell CP. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. medRxiv. 2020.

[18] 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. New England Journal of Medicine. 2020.

[19] Du Z, Wang L, Cauchemez S, Xu X, Wang X, Cowling BJ, et al. Risk for Transportation of 2019 Novel Coronavirus (COVID-19) from Wuhan to Cities in China. medRxiv. 2020.

[20] Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (COVID-19). medRxiv. 2020.

[21] Liu T, Hu J, Kang M. Transmission dynamics of 2019 novel coronavirus (2019-nCoV). bioRxiv 2020; published online Jan 26. DOI.10(2020.01):25.919787.

[22] Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of Travel Medicine. 2020.

[23] Cao Z Zhang Q, Lu X et al. Estimating the effective reproduction number of the 2019-nCoV in China. medRxiv 2020.

[24] Jung S-m, Akhmetzhanov AR, Hayashi K, Linton NM, Yang Y, Yuan B, et al. Real-Time Estimation of the Risk of Death from Novel Coronavirus (COVID-19) Infection: Inference Using Exported Cases. Journal of Clinical Medicine. 2020;9(2):523.

[25] 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. International Journal of Infectious Diseases. 2020.

[26] Zhang KK, Xie L, Lawless L, Zhou H, Gao G, Xue C. Characterizing the transmission and identifying the control strategy for COVID-19 through epidemiological modeling. medRxiv. 2020.

[27] Majumder MS, Rivers C, Lofgren E, Fisman D. Estimation of MERS-coronavirus reproductive number and case fatality rate for the spring 2014 Saudi Arabia outbreak: insights from publicly available data. PLoS Currents. 2014;6.

[28] Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet. 2020.

[29] Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner NW, Ke R. The Novel Coronavirus, 2019-nCoV, is Highly Contagious and More Infectious Than Initially Estimated. arXiv preprint arXiv:200203268. 2020.

[30] Park SW, Champredon D, Earn DJ, Li M, Weitz JS, Grenfell BT, et al. Reconciling early-outbreak preliminary estimates of the basic reproductive number and its uncertainty: a new framework and applications to the novel coronavirus (2019-nCoV) outbreak. medRxiv. 2020.

[31] Muniz-Rodriguez K, Chowell G, Cheung C-H, Jia D, Lai P-Y, Lee Y, et al. Epidemic doubling time of the COVID-19 epidemic by Chinese province. medRxiv. 2020.

[32] Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D. Estimation of the reproductive number of Novel Coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. International Journal of Infectious Diseases. 2020]

[33] Davide Bonin. Weather Data for COVID-19 Data Analysis, 2020. https://www.kaggle.com/davidbnn92/weather-data-for-covid19-data-analysis

[34] NOAA. Global Surface Summary of the Day, 2020. https://www.kaggle.com/noaa/gsod

[35] Wikipedia. D'Agostino's K-squared test, 2019. https://en.wikipedia.org/wiki/D%27Agostino%27s_K-squared_test

[36] Geopy. Geopy Documentation, 2020. https://geopy.readthedocs.io/en/stable/

[37] ABC News. China moves to block new virus flare-up on Russian border, 2020. https://abcnews.go.com/International/wireStory/china-moves-block-virus-flare-russian-border-70134699

Appendices

A Basic Reproduction Number Data Set

Study	R_0	95% CI Lower Bound	95% CI Upper Bound	Study Date
Lai et al. [9]	2.60	2.10	5.10	01/01/2020
Biao [10]	6.47	5.71	7.23	01/10/2020
WHO [11]	1.95	1.40	2.50	01/18/2020
Imai [12]	2.50	1.50	3.50	01/18/2020
Riou & Althaus [13]	2.20	1.40	3.80	01/18/2020
Zhou et al. [26]	5.50	5.30	5.80	01/20/2020
Shao et al. [15]	3.32	3.25	3.4	01/22/2020
Shen et al. [16]	6.49	6.31	6.66	01/22/2020
Shen et al. [16]	4.71	4.50	4.92	01/22/2020
Read et al. [17]	3.11	2.39	4.13	01/22/2020
Li et al. [18]	2.20	1.40	3.90	01/22/2020
Du et al. [19]	1.90	1.47	2.59	01/23/2020
Li et al. [20]	2.23	1.77	3.00	01/23/2020
Liu et al. [21]	2.90	2.32	3.63	01/23/2020
Liu et al. [21]	2.92	2.28	3.67	01/23/2020
Cao et al. [23]	4.08	3.37	4.77	01/23/2020
Jung et al. [24]	2.10	2.00	2.20	01/24/2020
Zhao et al. [25]	2.24	1.96	2.55	01/24/2020
Jung et al. [24]	3.20	2.70	3.70	01/24/2020
Zhou et al. [14]	3.58	2.89	4.39	01/24/2020
Majumder et al. [27]	2.55	2.00	3.10	01/26/2020
Joseph et al. [28]	2.68	2.47	2.86	01/28/2020
Wu et al. [28]	2.68	2.47	2.86	01/28/2020
Sanche et al. [29]	4.70	2.80	7.60	01/30/2020
Sanche et al. [29]	5.70	3.80	8.90	01/30/2020
Sanche et al. [29]	6.30	3.30	11.30	01/30/2020
Park et al. [30]	3.10	2.10	5.70	01/31/2020
Zhao et al. [25]	2.12	2.04	2.18	02/02/2020
Liu et al. [22]	4.50	4.40	4.60	02/07/2020
Liu et al. [22]	4.40	4.30	4.60	02/07/2020
Muniz-Rodriguez et al. [31]	3.30	3.10	4.20	02/09/2020
Zhang et al. [32]	2.28	2.06	2.52	02/16/2020

B Change in exponential growth rates per government response per country

	School closing	Workplace closing	Cancel public events	Restrictions on internal movement	International travel controls
Afghanistan	0.50696	-0.17778	1.62362	0.57132	1.81529
Argentina	0.52173	0.52173	0.42232	0.31909	0.40639
Australia	-0.00345	-0.40363	0.76472	0.42903	-0.8359
Austria	0.07461	0.73536	0.87418	0.43529	1.0899
Bahrain	-2.14402	0.06473	-0.12046	-0.09103	-0.13105
Belgium	0.21074	0.46051	0.08136	0.31389	1.42048
Bosnia and Herzegovina	2.60686	0.50059	2.60686	0.36335	2.45516
Brazil	1.10918	0.8603	1.10918	1.00916	0.8603
Burkina Faso	3.02175	-0.65322	3.00639	-0.07994	3.00639
Chile	0.82585	0.91033	0.91033	-0.1138	-0.02275
China	0.4016	0.4016	2.00687	1.36881	-0.11301
Costa Rica	-0.63404	0.14209	0.14209	0.25823	0.22771
Cuba	0.17028	-0.0697	0.56257	0.56257	0.56257
Czech Republic	0.59642	0.32272	0.59642	-0.14271	1.204
Denmark	-1.60376	-1.25707	1.45398	3.04297	3.04297
Ecuador	2.4314	1.64196	2.4314	1.64196	-0.74402
Egypt	-0.24928	0.09	-0.26807	-0.24928	-0.24928
Estonia	-1.23382	0.10951	1.61521	1.61521	1.61521
Ethiopia	-0.42804	0.31713	-0.42804	0.22953	-0.42804
Finland	-0.22952	0.65602	0.65602	-0.56953	-0.22952
Germany	3.14258	-0.45059	1.63921	0.33767	1.62796
Greece	1.78054	-0.19851	1.47064	-0.11128	-0.73327
Hungary	0.73063	0.2558	0.73063	0.73063	0.83612
Malaysia	0.8307	-0.41946	0.50417	-0.41946	-0.09073
Morocco	-0.02346	-0.02346	0.30252	1.82895	1.82895
Netherlands	0.1398	0.1398	0.055	0.1118	0.11379
Norway	-0.6077	0.49959	-0.6077	-0.58325	-0.18925
Paraguay	0.00084	0.06975	0.00084	0.4385	0.4385
Philippines	-0.32385	0.31384	-0.74194	0.00407	0.94446
Poland	0.80602	0.4177	0.90913	0.80602	1.25071
Portugal	1.22614	0.47725	1.28171	0.0513	1.07016
Saudi Arabia	1.15629	0.32174	-0.2241	1.61252	-0.2241
Slovenia	-0.94292	-0.32574	0.99078	-0.37561	0.52082
South Korea	-0.70841	-0.2488	-0.02894	-0.38268	-0.70841
Spain	1.36052	1.36052	0.41033	1.36052	0.41033
Switzerland	0.25724	0.01562	1.34238	-0.41629	0.25724
Thailand	1.64611	0.51174	1.3109	0.86188	0.56053
Tunisia	2.20018	0.55608	1.95258	0.73087	1.95258
Ukraine	1.81659	2.51803	1.81659	1.81659	1.81659
United States	1.78531	0.67989	1.27797	1.28692	-0.49401
Vietnam	-0.51594	-0.16351	-0.66681	-0.47717	1.15643