Monte Carlo Projection of COVID-19 Time Evolution for Italy and the United States Using Logistic Regression

Caroline Thomas, Spencer Kuhn, George Laird

Abstract

This project seeks to build off the paper "Mathematical prediction of the time evolution of the COVID-19 pandemic in Italy by a Gauss error function and Monte Carlo simulations" by Ignazio Ciufolini and Antonio Paolozzi in The European Physical Journal Plus [1]. In the paper, Ciufolini and Paolozzi use a Gauss Error function to project the "flex points," or inflection points, of the graphs of cumulative positive COVID-19 cases over time for Italy and China, therefore predicting the point at which the rate of added cases begins to decelerate. However, given the error present in both unreported cases and delayed reports, a single projection is unreliable. Therefore, Ciufolini and Paolozzi used a more robust method utilizing Monte Carlo simulations. They multiplied the daily case totals by a vector of draws from a normal distribution with a mean of 1 and a standard deviation of 0.1, estimating that measurement discrepancies represented somewhere in the area of 10% of each daily case total. This was repeated 150 times with 150 random vectors. Then, Gauss Error regressions were fit to each adjusted set of case totals, and the flex points of each were calculated.

Our simulation replicates the portion of Ciufolini and Paolozzi's paper in Italy, however, we employ a logistic regression rather than a Gauss Error regression, and we then use the same projection methods for the United States. In the section of our study replicating the procedures Ciufolini and Paolozzi used to find the average flex date for Italian projections, we reached a similar result with a date of March 23rd, just two days before the Ciufolini and Paolozzi prediction of March 25th. Using percentile intervals, we found that 95% of the flex dates fell within a range between March 19th and April 3rd. For the United States extension, our average flex date was evaluated at April 7th, with 95% of flex dates falling between April 3rd and April 12th. In Italy, the localized maximum for daily case totals during the first spike in cases was reported on March 22nd. In The United States, two localized maxima were reported on April 6th and April 11th. Localized maxima for the first case spikes are estimators for the flex points for the true time evolution of the virus. When compared to the dates of administered lockdowns in Italy and the United States, it becomes clear that the Italian response, which came earlier compared to our estimated flex points, more effectively reduced daily case counts in a manner consistent with our projections and a logistic model in general, wheras the United States did not reduce case counts in a manner consistent with our projected logistic models.

Introduction

The rapid proliferation of COVID-19 throughout the world sent countless researchers on a mission to project how the deadly virus would spread and how dangerous it might become. The earliest nations to suffer large outbreaks were China, Italy, and Iran, the first two of which are the subject of the paper "Mathematical prediction of the time evolution of the COVID-19 pandemic in Italy and China by a Gauss error function and Monte Carlo simulations" by Ignazio Ciufolini and Antonio Paolozzi in The European Physical Journal Plus. This study was conducted in March 2020 and published in April 2020 during the first wave of the pandemic. Noticing the need for modeling in order to assess the necessity of public health intervention by these nations, Ciufolini and Paolozzi took early case counts from Italy and China and used a Gauss Error regression to determine when the first spike in cases might be over. For the sake of time and brevity, we elected to focus on only reproducing their results for Italy. The following shows the rise in daily COVID-19 cases for Italy over time (Figure 1) using more recent case data, as well as the numbers from between February 15th and March 29th (Figure 2), which are the dates used in the original study, according to data provided by the

World Health Organization (WHO). The red line represents cumulative positive COVID-19 cases and the blue line indicates daily new positive COVID-19 cases.

Figure 1: Overview of COVID−19 Cases in Italy

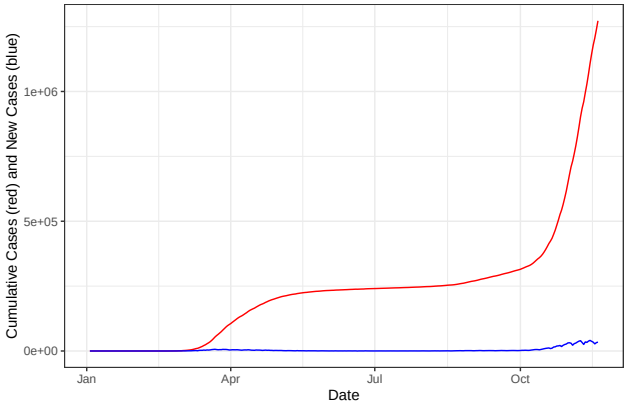
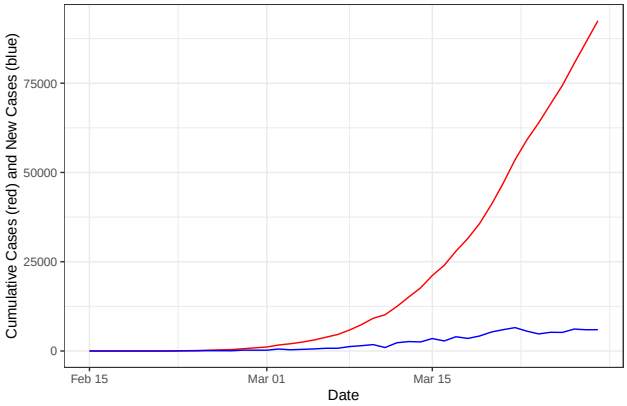


Figure 2: COVID−19 Cases in Italy Feb 15th to March 29th



In the United States, a similarly dramatic initial rise in cases was observed, although many argue that the United States has not in fact emerged from it's first case spike. The overall COVID-19 time evolution of the United States is plotted below (Figure 3), as well as the cumulative and daily case totals between February 1st and April 11th (Figure 4). These dates were selected because they occur prior to widespread lockdown procedures being implemented in the US so the data can give us an understanding of the initial transmission of the virus and to allow us to evaluate how lockdown procedures affected disease transmission trends in comparison to what would be expected given our model.

Figure 3: Overview of COVID−19 Cases in the United States

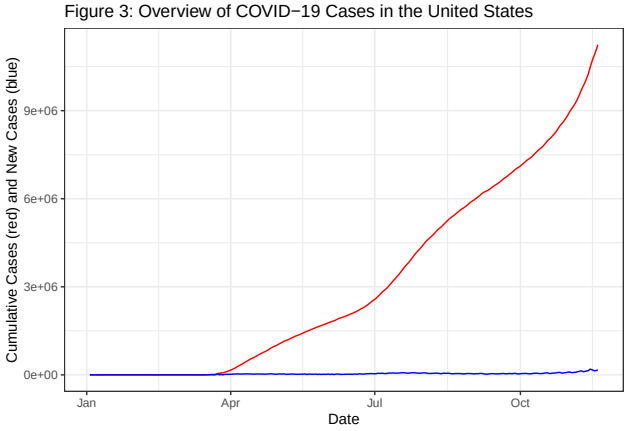
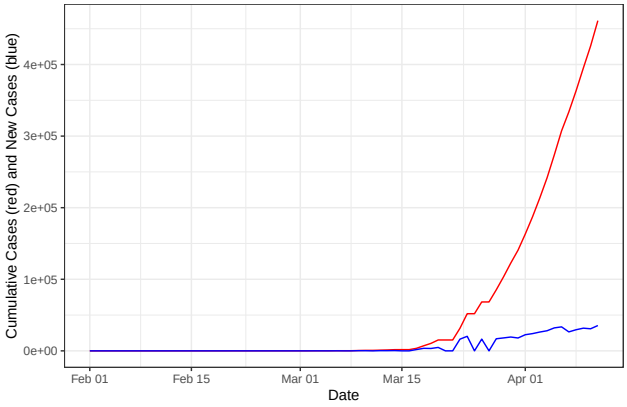


Figure 4: COVID−19 Cases in US Feb 1st to April 11th



However, case totals are frequently inaccurate due to delayed reporting, false positive or negative test results, and the appreciable number of individuals who had a symptomatic or asymptomatic case of the virus but did not get tested. Further, testing procedures have rapidly fluctuated due to availability of tests, advancements in testing methods, and public health testing recommendations. Thus, projected case totals need to be better validated, and some measure of error is necessary in order to evaluate the true range of possibilities. Monte Carlo simulations can be used to test hypothetical sampling error and the possible projections that may result from slight changes to daily case totals.

Methods

Using RStudio and GitHub, we drafted code to first replicate the Italian predictions made by Ciufolini and Paolozzi. This involved downloading the repository of daily COVID-19 cases across all countries made available in a csv file by the World Health Organization [5]. The Italian cases between February 15th and March 29th were extracted and then run through a nonlinear least squares regression with a logistic equation as the model. The model equation is below:

$$\frac{a}{1+e^{-(b+cx)}}$$

Wherein a is the upper asymptote, or the maximum number cumulative cases, and b and c denote intercept and slope parameters. Initial parameter estimates were made using an upper asymptote estimate of 100,000 and then finding linear regression parameters of the log odds of cumulative cases per day. Our estimate of 100,000 comes from the understanding that the flex date, based on a visual analysis, may be within the range of dates we selected for our regression data, and that 50,000 cases may roughly be the point halfway towards the final case total at the end of the first spike. The nonlinear least squares (NLS) regression was then used to find final cumulative case projections (parameter "a") and flex point projections (the opposite of parameter "b" divided by parameter "c"). NLS iterates through various adjustments to the model parameters in order to optimize the model fit. This method of finding the projection parameters and flex points is adapted from a "Marine Global Change Ecology" program at the University of Massachusetts Amherst [2]. The primary reason we elected to use NLS with a logistic model, rather than a Gauss Error function, has to do with the difficulty of using R to approximate parameters in the Gauss error function, whereas there are more methods in place for a simple logistic regression (Ciufolini and Paolozzi conducted their regression coding in Python). The original paper provided insufficient detail regarding their methodology for us to reproduce their modeling function in R. It was however noted in the original paper that logistic regression was tested and produced similar results which led us to consider GLM or NLS methods. Our methods yield similar flex dates to the published study so there is not a large discrepancy between our results and the results we attempted to reproduce. The code below details our method for finding a single projection.

```
#gathering initial coefficient estimates for NLS
coef(lm(logit(italy_dat_early$Cumulative_cases/100000)~italy_dat_early$index))
```

```
##           (Intercept) italy_dat_early$index
##            -10.242468              0.293917
```

```
#running NLS algorithm
italy_log <- nls(italy_dat_early$Cumulative_cases~a/(1+exp(-(b+c*italy_dat_early$index))),
                 start=list(a=100000,b=-10.242468,c=0.293917),
                 data=italy_dat_early,trace=FALSE)

summary(italy_log)
```

```
##
## Formula: italy_dat_early$Cumulative_cases ~ a/(1 + exp(-(b + c * italy_dat_early$index)))
##
## Parameters:
##     Estimate Std. Error t value Pr(>|t|)
## a  1.247e+05  2.000e+03   62.36   <2e-16 ***
## b -7.296e+00  6.326e-02 -115.33   <2e-16 ***
## c  1.885e-01  2.435e-03   77.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 555.6 on 41 degrees of freedom
##
## Number of iterations to convergence: 5
## Achieved convergence tolerance: 6.026e-06
```

```
#calculating flex point based on coefficients
flex_point_it <- -coef(italy_log)[2]/coef(italy_log)[3]
```

```r
#creating a projection graph based on the model
projection_italy <- (coef(italy_log)[1])/(1 + exp(-(coef(italy_log)[2]+
                                                    coef(italy_log)[3]*1:100)))

#creating a vector of dates for the x-axis
proj_dates_italy <- data.frame(projection_italy, "date" = dates[44:143])
```

Once the procedure was completed for one data set, it was then adapted to the 150 Monte Carlo simulations. The original paper very clearly laid out their Monte Carlo procedure, although did not provide the reasoning behind their method. For the Monte Carlo simulations, an m by n matrix was created where m was the number of Monte Carlo simulations desired (150 in our case) and n was the number of days we wanted to simulate. Then, the matrix was filled with randomized draws from a normal distribution with a mean of 1 and a standard deviation of 0.1. Each column of the matrix was then multiplied by the daily case total with the corresponding index. Using a standard deviation of 0.10 allowed for 10% uncertainty to be incorporated into the simulated cumulative case data. Thus, 150 simulated flex dates were obtained using the NLS regression model described above and then plotted along with the mean of these dates. The Monte Carlo simulations allowed Ciufolini and Paolozzi to not only establish a more reliable average flex date, but also determine some measure of error in the distribution of flex dates so that statistical inference could be used to better validate their findings. In particular, 95% confidence intervals using t-distributed quantiles with 149 degrees of freedom were made to estimate the probable range within which the true flex date resides. The distribution of Monte Carlo simulated flex dates were also described using their 2.5th and 97.5th percentiles. The Monte Carlo simulation and inference algorithms are below:

```r
#writing a function to simulate Monte Carlo random vectors
sim <- function(m, n) {
  mat <- matrix(data = NA, nrow = m, ncol = n)

  for (i in 1:m){
    mat[i,] <- rnorm(n, sd = 0.1, mean = 1)
  }

  return(mat)
}

#using cum cases from Feb 15th to March 26th
cum_days <- italy_dat[44:84,6]

#creating 150 simulations for 41 days
simulations <- sim(150, 41)

#writing a function that multiplies the simulation vectors by the case count vector
mult <- function(x, y){
  new_mat <- matrix(data = NA, nrow = 150, ncol = length(x))
  for (i in 1:length(x)) {
    new_mat[,i] <- x[i]*y[,i]
  }
  return(new_mat)
}

#multiplying the cumulative days vector by each simulation
test <- mult(cum_days, simulations)

mc_mat <- as.data.frame(test)
```

```r
#running logistic regression for 150 simulations and get flex date for each

flex_mc <- rep(NA, 150)

for (i in 1:150) {
  row_data <- t(mc_mat[i,])
  index <- seq(1:length(row_data))
  coefs <- coef(lm(logit(row_data/100000)~index))
  italy_log <- nls(row_data~a/(1+exp(-(b+c*index))),
                   start=list(a=100000,b=coefs[1],c=coefs[2]),trace=FALSE)
  flex_mc[i] <- -coef(italy_log)[2]/coef(italy_log)[3]
}
```

For the US projection, we used case counts from between February 1st and April 11th in order to obtain a sense of the projected case totals prior to many of the measures taken by local and federal governments to flatten the curve. Initial asymptote estimates were set at 700,000 because a visual analysis of the cumulative case totals for the United States suggests, like Italy, that our data may include the flex date and that 350,000 cases may be the point roughly halfway towards the final case number at the end of the first spike. The flex date algorithm was otherwise identical, with an average date calculated along with its corresponding 95% confidence intervals (a tradition t-quantile confidence interval and a percentile confidence interval).

Results

Figures 5 and 6 show our fitted regression lines with raw data projections for Italy and the United States. The models seem to fit the data well and provide projections for the progression of cumulative cases into the month of May, following a logistic curve. Our fitted regression for Italy gives us a flex date of approximately 39 days after February 15th, or March 24th. This result is very close to the projected flex date in the original paper which was March 25th. Our fitted regression for the United States gives us a flex date of approximately 66 days after February 1st, or April 6th.

Figure 5: Cumulative Cases with Fitted Regression Line for Italy

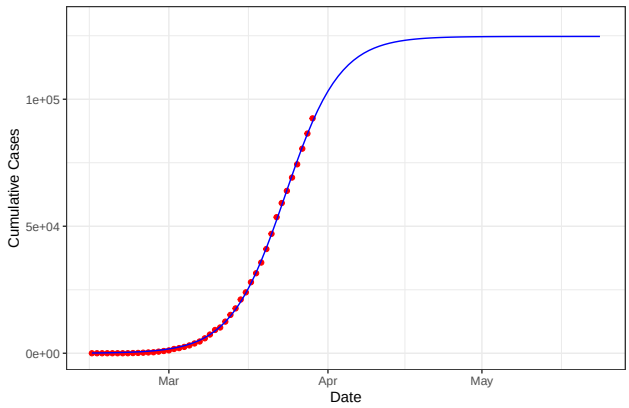
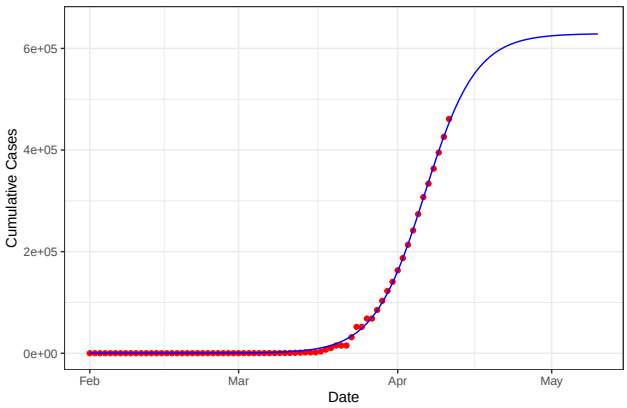
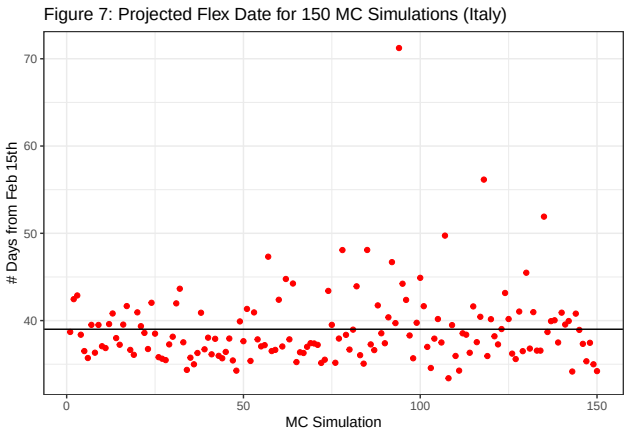
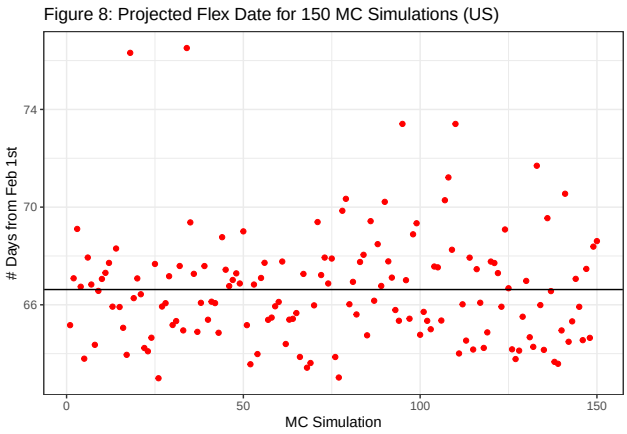


Figure 6: Cumulative Cases with Fitted Regression Line for US



Figures 7 and 8 below show the 150 Monte Carlo simulated flex points of each projection for Italy and the United States with a horizontal line drawn to represent the mean flex date. Flex date is shown on the y-axis and simulation number (1-150) is on the x-axis.





Corresponding to these plots, we gathered the mean, approximate flex date, and standard deviation for Italian and US flex points as well as a 95% confidence interval for each country. Additionally, we computed the 2.5th and 97.5th percentiles for each set of flex dates to create 95% percentile confidence intervals. These results are summarized in Table 1 below. Our Monte Carlo simulations give us a projected flex date for Italy of March 23rd and a projected flex date for the United States of April 7th. These dates are both similar to (within one day of) the flex dates projected by our regression models, however, the Monte Carlo method provides us with a better measure of uncertainty in these projections and allows us to create intervals of probable flex dates. The computed 95% confidence intervals using t-quantiles are relatively narrow, however, the percentile confidence intervals provide a much wider range of dates. The percentile intervals likely provide a more accurate understanding of the uncertainty in the flex date estimates. For Italy, our percentile confidence interval captures a range of flex dates from approximately March 19th to April 3rd. For the United States, our percentile confidence interval captures a range of flex dates from approximately April 3rd to April 12th.

Table 1: Summary of Monte Carlo Results

	Mean # of Days	Approx. Flex Date	SD of # of Days	95% CI	2.5th percentile	97.5th percentile
Italy	39.0	March 23rd	4.4	(39.7, 38.3)	34.3	48.5
United States	66.6	April 7th	2.3	(67, 66.3)	63.6	72.2

Discussion

Over the last year, COVID-19 has significantly altered the lives of almost every person on the planet, and the virus has killed over 1.5 million globally. As world leaders struggle to contain the spread of the virus, it's imperative to establish which public health plans are effective and which ones are not. Those in charge of handling this pandemic must be held accountable for their response to one of the biggest challenges of the twenty-first century. While there is often little a government can do about preventing the virus from entering their borders entirely, there have been numerous examples of effective national programs that "flattened the curve" once the first COVID cases were reported. In this paper we establish whether Italy and the United States managed to intervene and reduce daily case totals below mathematical projections.

There are some notable shortcomings in our data to be considered alongside the results. To begin, NLS regression uses an optimization algorithm that differs from a traditional least squares method. In NLS, numerous local maxima may be present, meaning the algorithm may settle on a curve fit that isn't necessarily the most appropriate. Furthermore, NLS will terminate if too many iterations are required to find a proper fit. We ran into this issue with a few randomization seeds, finding that several data points had to be removed

because no NLS coefficients could be found. This was combated by adjusting the initial parameters set for the algorithm. Moreover, initial asymptote estimates were appreciably lower than the ultimately reported cumulative case totals for Italy and the United States. Thus, different asymptote estimates may have yielded different flex date estimates and model coefficients based on the way the NLS algorithm operates.

In terms of replicating the experiments of Cuifolini and Paolozzi, we were able to reach an average flex date very close to their own for Italy. Any discrepancies are likely due to the NLS method as opposed to the original method using a Gauss Error function. The Monte Carlo simulation method allowed us to better assess error and uncertainty in the reported numbers and the projection that resulted. In the extension of this method to the United States, we ran into some issues with the initial asymptote estimate. the NLS algorithm ran into issues with our first estimate of 500,000, so we changed the estimate to 700,000 in order for NLS to run successfully for all Monte Carlo simulation vectors.

For both countries, our calculated average flex dates fell very close to each localized maxima for daily case totals during the first spike in cases. As seen in figure 9 and figure 10 below, the localized maximum number of daily cases, an estimate for the true date of the flex point, was within our confidence range for both Italy and the United States. In Italy, two local maxima are seen around March 22nd and March 27th, meaning the true flex date is likely somewhere in between. This falls right around our mean flex date of March 23rd. In the United States, two local maxima are seen around April 6th and April 11th. Our mean flex date for the United States was calculated to be April 7th.

Figure 9: Italy daily cases between March 2nd and April 1st

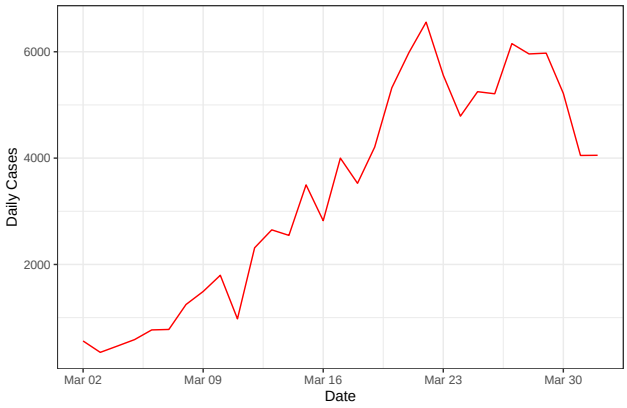
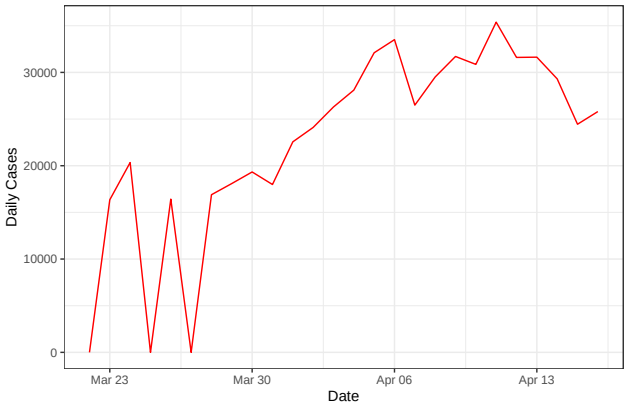


Figure 10: US daily cases between March 22nd and April 16th



Despite our means being accurate to the observed data, our Monte Carlo simulations yield percentile confidence intervals that are appreciably wide (March 19th - April 3rd in Italy and April 3rd - April 12th in the US), and it is clear that small changes to the daily case data produce very different time evolutions and flex dates for the progression of COVID-19. It is not clear what type of confidence intervals were used by Ciufolini and Paolozzi, so we will not compare ours with theirs. We can however speak to the importance of the flex date, as it represents only the halfway point of a spike in cases, meaning a later flex date incurs an even longer period over which the first spike resolves.

The projections made for Italy reflect numbers for a nation that had instituted a nationwide lockdown nearly two weeks before the estimated flex date, when case counts were less than 10% of the estimated maximum [4]. In the United States, only 42 states had issued stay at home orders by the time of the estimated flex date, and by the time more than half of the US states had issued stay at home orders, case counts were over 15% of the estimated maximum. Furthermore, only nine states had issued stay at home orders by two weeks before our estimated flex date [3]. Thus, our projections for Italy represent those of a nation already intervening heavily with public health programs, whereas our projections for the United States represent those of a nation still struggling to commit to the kind of drastic public health measures and restrictions needed to slow the spread of COVID. The consequences of this delayed response are seen in Italy's impressive decline in daily cases between June and August, whereas the United States continues to see high case totals throughout the summer as shown in Figures 1 and 3 above. Many projections, ours included, grossly underestimate the actual cumulative case counts in the United States, largely because stay at home orders were either not strict

enough or relaxed too soon, in our assessment, to facilitate a time evolution that can be modeled by logistic regression. Further research could be done to investigate other modeling methods using current cumulative case data to perhaps find a more suitable method for modeling the progression of COVID-19 across the world.

MLA Works Cited:

[1] Ciufolini, Ignazio and Paolozzi, Antonio. "Mathematical prediction of the time evolution of the COVID-19 pandemic in Italy by a Gauss error function and Monte Carlo simulations." The European Physical Journal Plus. vol. 135, no. 355, 2020.

[2] Cheng, Brian. "Modeling Logistic Growth Data in R." Marine Global Change Ecology. University of Massachusetts Amherst, 7 May 2014. https://bscheng.com/2014/05/07/modeling-logistic-growth-data-in-r/. Accessed 6 December 2020.

[3] Fernandez, Marisa. "Timeline: How the U.S. fell behind on the coronavirus." Axios. Axios Media, 10 April 2020. https://www.axios.com/coronavirus-timeline-trump-administration-testing-c0858c03-5679-410b-baa4-dba048956bbf.html. Accessed 8 December 2020.

[4] Lawler, Dave. "Timeline: How Italy's coronavirus crisis became the world's deadliest." Axios. Axios Media, 24 March 2020. https://www.axios.com/italy-coronavirus-timeline-lockdown-deaths-cases-2adb0fc7-6ab5-4b7c-9a55-bc6897494dc6.html. Accessed 8 December 2020.

[5] "WHO Coronavirus Disease (COVID-19) Dashboard." World Health Organization. World Health Organization, 8 December 2020. https://covid19.who.int/table. Accessed 8 December 2020.