Math 2820L-01 Statistics Lab

COVID-19 Geographical Analysis Report

Group members: Terry Luo, Raahul Natarrajan

Introduction: Our group attempted to analyze time series data from the recent COVID-19 outbreak. This disease has infected thousands of people worldwide and has forced the entire world to go under quarantine. The infectiousness of this disease is remarkably high. As of April 2, 2020, there are over 2.2 million cases worldwide. Thus, the need to learn more about this virus, its effects, and behavior is paramount in order prevent further outbreaks and infection. Our analysis will focus on how the geographical location of a local outbreak affects the rate of spread for the virus. Our primary source of data is the COVID-19 public data repository by Johns Hopkins University, which keeps track of daily number of confirmed cases for a variety of locations. We limited this analysis to the US only, for reasons stated later

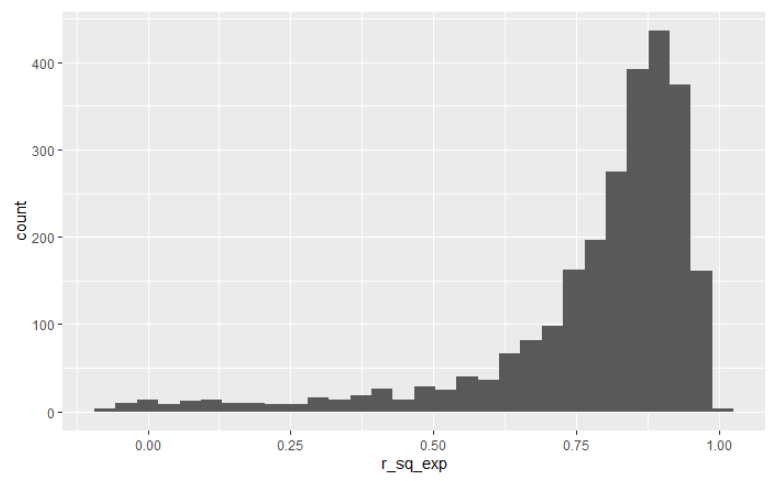
First, we performed regression on the number of confirmed cases to an exponential model. For each location, we extracted the number of cases starting from the day that location reported its first case. In an exponential equation P=Aert, the growth rate which we will call k, is encapsulated by the coefficient r. This is our measure of rate of growth, assuming that the number of cases in an area grows exponentially. We confirmed the legitimacy of this approach by finding the R-squared values of each regression. As seen in Figure 1, most R^2 values are very high. 

Figure 1 R-squared values for exponential regressions

At this point, we also attempted to make a conclusion about growth rate (k) and location. Figures 2 and 3 compare growth rate versus latitude and longitude.

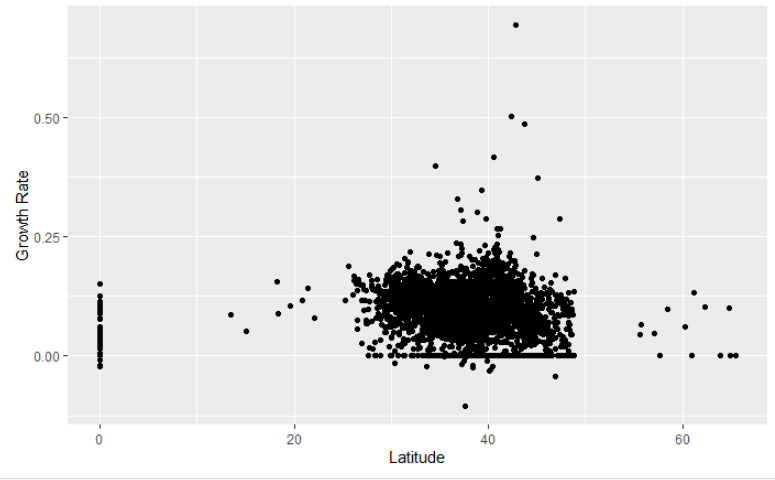


Figure 2 Growth Rate vs Latitude

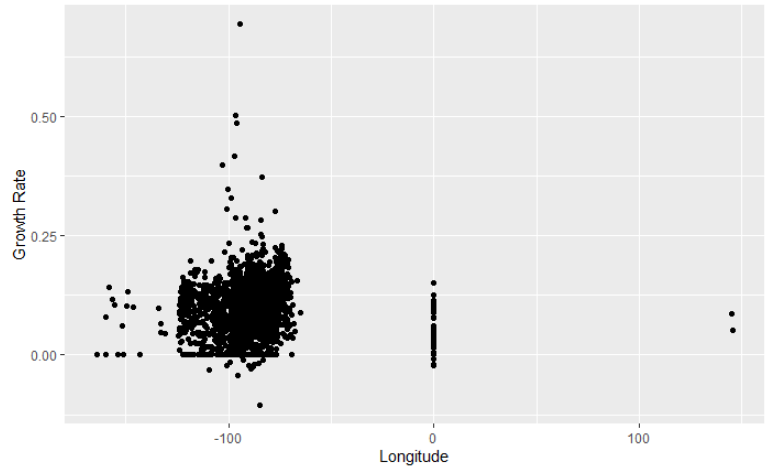


Figure 3 Growth Rate vs Longitude

These two figure demonstrate that by location coordinates alone, the rate of coronavirus spread is fairy uniform throughout the US. This is summarized in Figure 3, which isolates the data to the continental US.

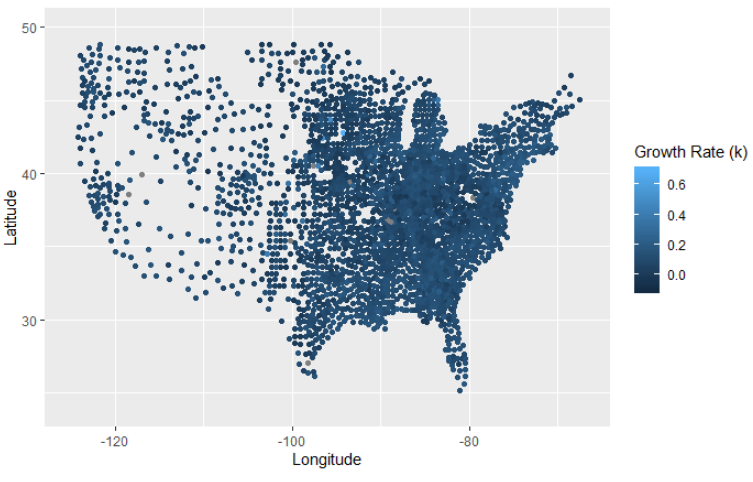


Figure 4 Growth Rate across the US

Looking back at Figure 1, although regression of most location’s case count to an exponential model was very good, there were still a few locations where the exponential model failed. In these cases, we theorized that the number of cases had reached an asymptote, and a regression to the logistic growth model would be better. By comparing the r squared values from regressions to both the exponential and logistic growth models, we hoped to classify whether the virus at each location was still spreading rampantly or had been contained. If the exponential model fit the data better, then we could conclude the virus was still spreading. If the logistic growth model fit the data better, then we could conclude the virus had been contained.

Unfortunately, with our current skills in R and existing libraries, we were not able to fit the data to a logistic model. The reason for this is that nonlinear models cannot extrapolate whether the data looks to be logistic or not. Even if the number of cases has reached an inflection point, which signals the possibility of an asymptote, if the asymptote has not occurred yet, the regression can’t be completed.