Coronavirus:

An analysis of Cases, Deaths and Recoveries

Lauren Campbell, Jill Mercer, Rachel Sabatino

CPSC 685  
*Slippery Rock University*Slippery Rock, PA, USA  
{lxc1050, jmm1037, rbs1004}@sru.edu

*Abstract*— Monitoring and predicting trends in disease plays an important role in how and where people will become infected. As the novel coronavirus (COVID-19) becomes a pandemic looking at how the virus will continue to spread can be accomplished through several means. Taking a look at regression analysis and clustering techniques we aim to answer some of these key questions.

Keywords— disease, COVID-19, SARS, coronavirus, SIR, cluster analysis, outlier detection, clustering, trends

# Introduction

On December 31, 2019 Wuhan, China reported the first case of a virus that would become a worldwide pandemic. Coronavirus has rapidly spread across the globe and continues to introduce a great deal of unknowns. With these unknowns come questions and concerns from the population. Medical professionals are learning everything they can to treat the disease and prevent further contamination, government officials are working to create legislation, rules, and regulations for travel bans, vaccine development, and financial help for those who will be affected.

Daily there are updates of new social gatherings being canceled. Colleges and Universities, The National Basketball Association, National Hockey League and Professional Golf Association are just a few to have postponed their year due to the outbreak.  The virus has the possibility to affect everyone, so there is great interest to learn more about the virus in order to better treat, prevent, and protect the population. For these reasons, this topic is of interest to our group.

# Related Work

Although this strain of Coronavirus, often referred to as COVID-19, is new, coronaviruses are not a new classification. These viruses were first discovered in the early 1930s in chickens and later discovered in humans in the 1960s. The term coronavirus is known as a range of animal species that cause respiratory, gastrointestinal, neurological, and systemic diseases. SARS has since been determined to be caused by a novel coronavirus and two other viruses names NL63 and HKU-1 are also associated with the disease (Peris). While a different type of influenza, the 1918 influenza pandemic can lead to useful perspective for today’s COVID-19. After 1918 is better understood: all three pandemic viruses since 1918 contain gene segments derived from the 1918 “founder” virus. Consequently, the past 94 years can be considered to constitute a single “pandemic era,” with the founding 1918 pandemic virus leading to emergence of pandemic progeny viruses in 1957, 1968, and 2009. The 1918 pandemic became more deadly after a mutation when the summer ended. In order to better predict the emergence of future pandemic viruses, structure/function correlates are critically needed to understand the biological implications of mutation patterns associated with new-host adaptation. [1]

Previous work has also been done in modeling the spread of infectious diseases such as SARS and the Ebola outbreaks.  One such approach is to focus on a clustering method that is not nonparametric, an example being hierarchical clustering methods.  These methods of clustering tend to eliminate biological factors that can be important for predicting those most likely to become infected and thus leading to a more inaccurate data model.  Research was done on tree algorithms to incorporate the differences that disease has on individuals more at risk for catching and ultimately dying from the spread of these diseases. [2]

As this is a developing issue, there is not a great level of knowledge as to what is causing this strain to be so infectious.  What is known so far about the disease, is heavily based on previous strains of the virus. It is known that common symptoms include fever, cough, and shortness of breath. Symptoms generally arise 2-14 days after initial exposure and can infect others through the air.

# Methodology

## Dataset

The data that will be used is compiled by Johns Hopkins School of Public Health and is housed on the Humanitarian Data Exchange site. There are three files provided one of which contains Confirmed Cases, and the others containing deaths and recoveries. Each file follows the same format with the following fields:

Providence/State

County/Region

Lat

Long

There are fields for each day in which data is available, starting with January 22, 2020. As the number of confirmed cases, death and recovery numbers continue to increase, the most recent data to date will be utilized.  The Johns Hopkins dataset is updated daily with information from the Center for Disease Control and Prevention (CDC) organizations worldwide. [3]

## Tools

The data will be analyzed through the use of R and by using the different data sets involving Infected, Recovered and Death rates through the date April 10, 2020. Due to the nature of the dataset, which is updated on a daily basis, we chose to pick the most recent dataset available at the time to work from.

## Preprocessing & Analysis Preparation

For our analysis we split the data into three pieces to allow detailed analysis for each variable (Infected, deceased and recovered). The first breakout of the data is those that were Infected with the COVID-19 virus. The data was downloaded from the John’s Hopkins site and saved into a csv file. Some formatting was completed on the raw file to create a merged column of Country and Region. Without this formatting the file had redundancy or lack of value if either County or Region was used alone. Once formatting was completed, the data was loaded into R, and a data frame was created. The first step in Preprocessing the data was to run summary statistics to assess the available fields, average values, and account for any data gaps. The data did not appear to have missing values or any qualities that would skew the analysis. Next, a box plot of the number of Infected cases was created to evaluate the range of the values to help determine how many clusters we will want to use for our experiment.

## Experiments

In order to create a better concept of where data was being collected, a graph of the locations was created based on the latitude and longitude. Only locations within the United States were used for the subject of concern.

Figure :First 48 States Datapoints

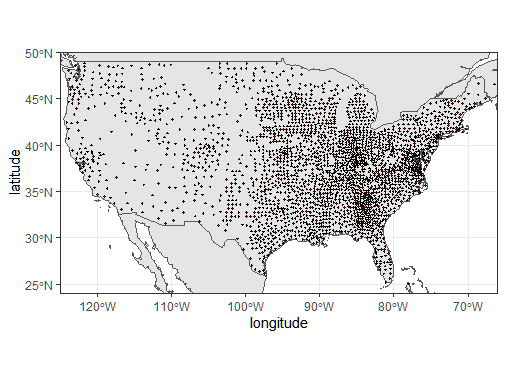


Figure : Hawaii Datapoints

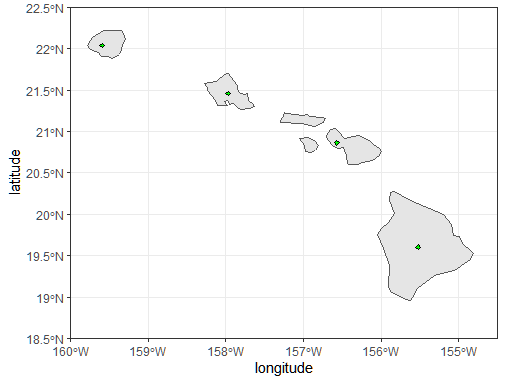
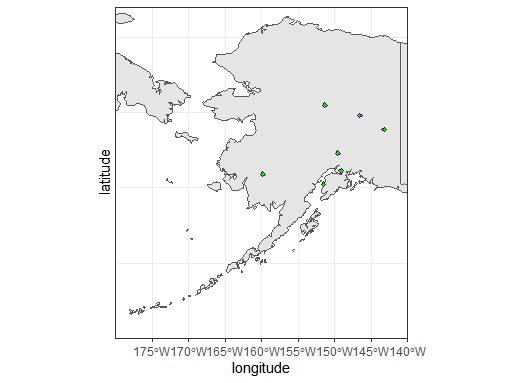


Figure :Alaska Datapoints



1. *Infected Analysis*

For the infected piece of our project, we want to try to identify any patterns in the data based upon volume of cases and geographic region. To do this we will use the k-means clustering algorithm. Based upon the research done with the data preprocessing and the wide range of infected cases in the data, an n = 5 was chosen for the number of clusters we will use. The K-Means algorithm gave the below results:

K-means clustering with 5 clusters of sizes 132, 225, 47, 10, 21498

Cluster means:

[,1]

1 67,803.5758

2 18,620.6133

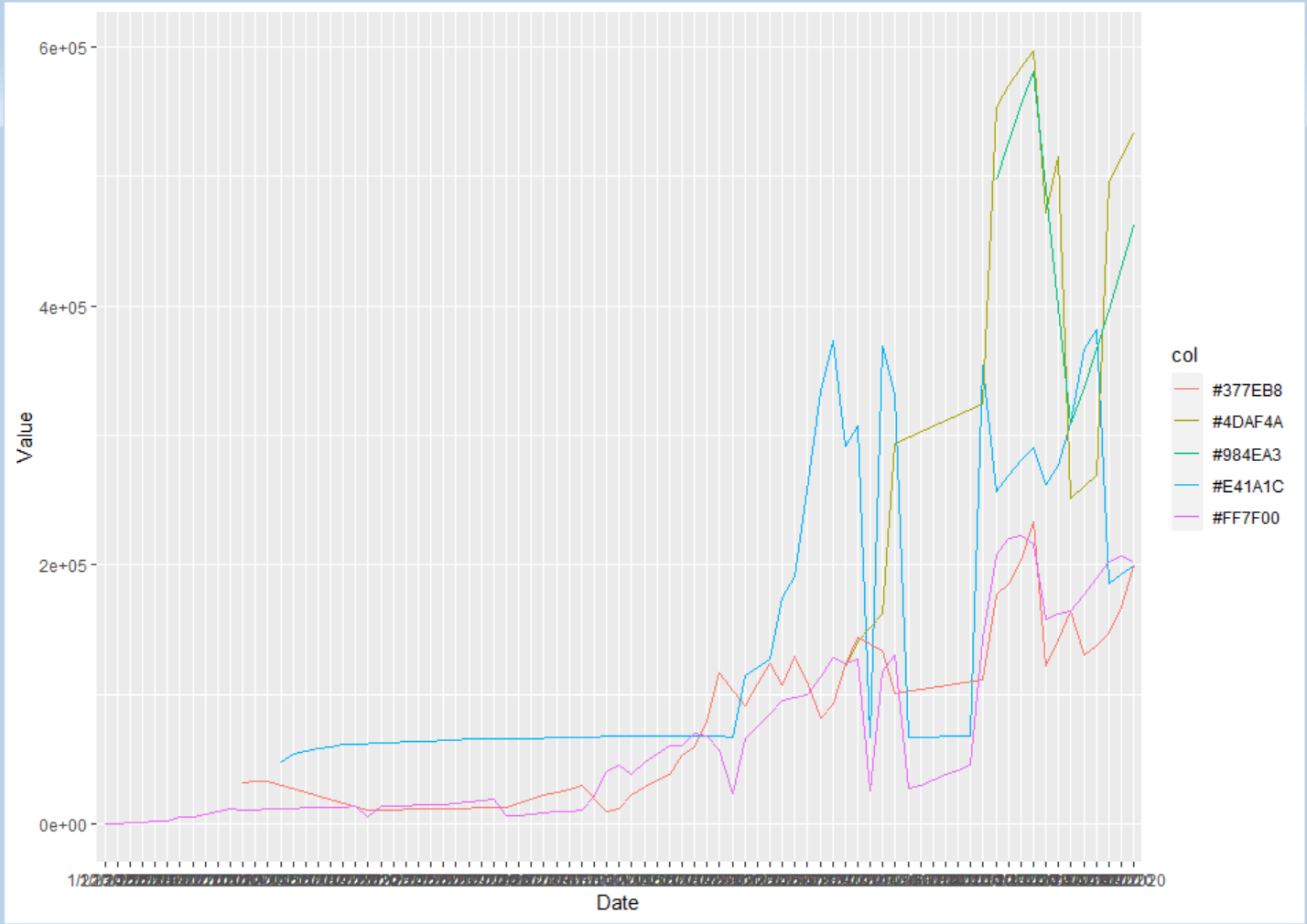
3 141,536.0213

4 445,816.7000

5 231.2348

The results indicate that most of the data points have a similar mean with a small percentage falling into the other four clusters. The cluster means have a significant range with the largest mean being 1,928 times larger than the smallest mean. Once the clusters are established, more analysis needs to be run to interpret the results. The original data frame was updated to add a column to represent which cluster each County\_Region belonged to. This allowed the total infected cases to be grouped by date, County\_Region, and then segmented by cluster. Using the ggplot2 command the volume of infected cases were graphed with the Date displaying on the x axis and the Volume on the Y axis with each cluster having a different line. The visual gave a representation of how the volume of each cluster moved over time. The visualization below shows each cluster as a different color and the increase in the number of Infected cases over time is the x axis. We can see that each cluster has a large spike starting around the same timeframe, however some of our clusters that have a higher volume appear to have a later date where the cases start to appear. With our knowledge of the virus we know that it hit different geographic regions at different points in time.

Figure :Infected Cluster Results



1. *Deceased Analysis*

The next set for analysis was the number of deceased. The focus on clustering for this portion was on latitude, longitude and the number of deaths reported. The goal for this is to run k-means based on geographic location and thus create clusters that represent areas with larger death totals than others. A k-means clustering analysis was utilized with Euclidean distance.

The k-means center number remained at 5 for consistency with the clustering of infected. The clusters produced the following results:

K-means clustering with 5 clusters of sizes 286, 18, 1230, 1083, 1

Cluster means:

Lat Long\_ Deaths

1 42.09527 -116.95229 3.4580420

2 39.99258 -83.12697 308.8333333

3 37.42962 -81.53769 3.9959350

4 37.66009 -94.45499 0.7137581

5 40.76727 -73.97153 5820.0000000

After k-means was preformed the cluster assignments were stored in the original data set in order to see geographically where these groupings were. A graph of the clusters in relation to geographic location was created:

Figure :Deceased Cluster Results

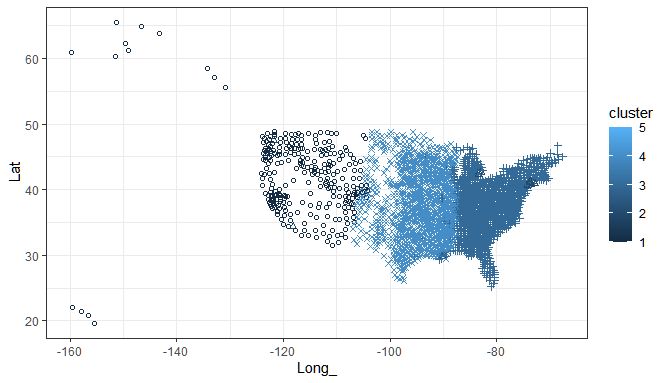
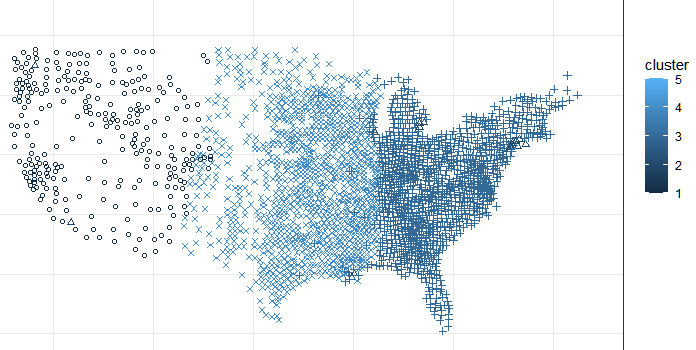


Figure : Deceased 48 States Cluster Results



The geographic locations appear to fall into East, Midwest and West of the United States with 2 other clusters forming in locations where death tolls are higher than average (Figure 5). As noted in the summary of the clusters, the largest mean death cluster is number 5 followed by 2.

The clustering grouped New York City as an outlier cluster at mean 5280, as the death toll as of 4/10/2020 was the highest in the country for a single data point. The following high death cluster (cluster 2) has 18 cities across the country that have higher than average death tolls. These cities included:

Bergen, Hudson, Union, Essex and Middlesex in New Jersey. Nassau, Rockland, Suffolk and Westchester in New York. Wayne, Oakland and Macomb in Michigan. Cook, Illinois, Fairfield Connecticut, Jefferson Louisiana, King Washington, Los Angeles California and Orleans Louisiana.

The remaining clusters grouped the East, Midwest and West into clusters 3, 4 and 1 respectively. This clustering layout reveals that the highest geographic area for death due to COVID-19 is on the East coast at a mean of roughly 4 people. It is important to note that the data used to create these clusters was from 4/10/20

# References

[1] Jeffery K. Taubenberger, David Baltimore, Peter C. Doherty, Howard Markel, David M. Morens, Robert G. Webster, Ian A. Wilson. *Reconstruction of the 1918 Influenza Virus: Unexpected Rewards from the Pas*t. Sep 2012, 3 (5) e00201-12; **DOI:** 10.1128/mBio.00201-12 <https://mbio.asm.org/content/3/5/e00201-12#sec-4>

[2] McCloskey RM, Poon AFY (2017) A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation. PLOS Computational Biology 13(11): e1005868.

[3] Novel Coronavirus (COVID-19) Cases.Humanitarian Data Exchange: John Hopkins School of Public Health. 13 March 2020.

<https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>