Winter 11-9-2022

Functional Data Analysis of COVID-19

Nichole L. Fluke

Nichole Lynn Fluke

Candidate

Mathematics and Statistics

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

James Degnan , Chairperson

Helen Wearing

Brent Wagner

Functional Data Analysis of COVID-19

by

Nichole Lynn Fluke

Bachelor of Science, Eberly College of Science

Major in Mathematics

Minor in Statistics

The Pennsylvania State University, 2014

Graduate Certificate, Applied Statistics

The Pennsylvania State University, World Campus 2017

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

Statistics

The University of New Mexico
Albuquerque, New Mexico

December, 2022

Dedication

I dedicate this work to my fiancé, Cody Gummo, who is always there for me and encouraging me to go further and also to my parents, who have been and are always there to help me, always encouraging me to reach for my dreams.

Acknowledgments

I would like to express my gratitude to my advisor, Professor James Degnan, for his guidance and support over the past three years. Our many sharing meetings about the research kept me ever searching for answers. I appreciate all the time that Professor Degnan spent discussing various research ideas with me. His valuable suggestions and comments while writing my thesis helped narrow the research and discussions.

I would like to thank my committee members Helen Wearing and Brent Wagner for their time reading and reviewing my thesis and participation in the defense of my thesis.

I would also like to extend my thanks to Mr. Jim Schaper, my high school math teacher from the State College Area School District, who inspired my interest in Mathematics which helped to guide me to my Undergraduate and Graduate Degrees and to my career in Mathematics and Statistics.

Functional Data Analysis of COVID-19

by

Nichole Lynn Fluke

B.S., Mathematics, The Pennsylvania State University, 2014

Graduate Certificate, Applied Statistics, The Pennsylvania State University, World Campus 2017

M.S., Statistics, University of New Mexico, 2022

Abstract of Thesis

This thesis deals with Functional Data Analysis (FDA) on COVID data. The Data involves counts for new COVID cases, hospitalized COVID patients, and new COVID deaths. The data used is for all the states and regions in the United States. The data starts in March 1st, 2020 and goes through March 31st, 2021. The FDA smooths the data and looks to see if there are similarities or differences between the states and regions in the data. The data also shows which states and regions stand out from the others and which ones are similar. Also shown is which seasons are the best or worst for COVID-19. The data mainly focuses on finding if the regions and states are the same or different for new COVID cases, hospitalized COVID patients, and new COVID deaths.

Functional Data Analysis of COVID-19

Nichole Fluke

November 08, 2022

Table of Contents

Table of Contents (Continued)

List of Figures

List of Figures (Continued)

List of Figures (Continued)

List of Figures (Continued)

List of Figures (Continued)

List of Tables

1.0 Introduction

As of August 2, 2021, there have been 35,768,924 Coronavirus cases and 629,380 deaths. [18] In December of 2019 a new virus emerged in Wuhan, China called COVID-19. [1]. It then started to slowly spread with more people being infected with COVID-19. As COVID-19 started to spread, it then went from country to country with more people becoming infected.

Anyone can get COVID-19 and experience any of the symptoms. [15] The CDC found that some people are more at risk of getting COVID-19 than others. Older adults or people having any severe or underlying medical conditions listed below in Table 1.0-1, are at a higher risk of getting more serious complications: [5]

Table 1.0-1, Severe / Underlying Medical Conditions

Conditions		
Cancer	Chronic kidney disease	Chronic lung diseases
Dementia	Diabetes	Down syndrome
Heart conditions	HIV	Immunocompromised state
Liver disease	Overweight or obese	Pregnant
Sickle cell disease	Smoking	Substance use disorders
Stroke	Solid organ transplant	Blood stem cell transplant

The most common way that COVID-19 spreads is from person to person. There are several ways that COVID-19 can spread from person to person: [11, 12]

- Droplets or aerosols - When a person that has COVID-19 coughs, sneezes, or talks, infected droplets or particles are suspended into the air. When this occurs, those infected particles can be inhaled by others who are within 6 feet from that person, and become infected.

- Airborne transmission - The virus that causes COVID-19 can live and stay in the air for up to 3 hours. If someone that has the virus exhales the infected air and someone else breathes the air in, they can then become infected with the virus.

- Surface transmission - If someone that has the virus coughs or sneezes onto a surface, then someone that doesn’t have the virus touches that surface, and then their eyes, nose or mouth, the uninfected person can become infected. The virus can stay on that surface for up to 2-3 days unless it is cleaned and disinfected.

- Fecal-oral - Researchers have found some evidence that if an infected person uses the bathroom and doesn’t wash their hands after, that they can then touch different surfaces or people and affect them as well.

This study uses Functional Data Analysis (FDA) to analyze different data sets. The study looks at positive cases, hospitalized patients, and deaths in the United States and compares the information across different states and regions.

2.0 Data

In this study there are three different types of data:

- new COVID cases,
- hospitalized COVID patients, and
- new COVID deaths.

New COVID cases show for that day the number of new COVID cases there are.[17] Hospitalized COVID patients is the reported patients currently hospitalized in an inpatient bed that have suspected or confirmed COVID-19.[6] New COVID deaths show for that day the number of new COVID deaths.[17]

This data might have some errors that cause accuracy issues. [10] The symptoms[15] might not show up right away, and there might be a delay in the testing[16] and reporting of having COVID. Some states/regions might update cases over time when more information is received and not right away. With the delay in new COVID cases, there might also be a delay and issues with the death data. [9] There are many steps involved in reporting death data, and this can cause a delay in the counts.

Figure 2.0-1, Data Timeline, contains the Data Day (0, 100, 200, 300 and 400), Season (Spring, Summer, Fall, and Winter), and corresponding Month of the data. The data starts March 1st, 2020, and goes to March 31th, 2021. [17] This data is from the Centers for Disease Control and Prevention (CDC). [17] Figure 2.0-1, Data Timeline, shows the layout of the data on graphs displayed in later sections.

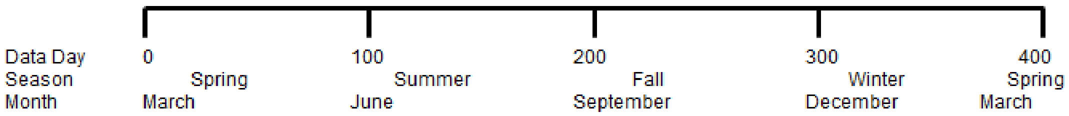

Figure 2.0-1, Data Timeline

In Figure 2.0-1, the data day starts at 0 and goes up to 400, this shows day 1 as March 1, 2020 and day 396 as March 31, 2021. Figure 2.0-1 has the layout of which month is it at which specific Data Day (0, 100, 200, 300, and 400) and the Season in which the Data Day is contained.

Within each of the three different types of data, the data is broken down by region and those regions states. [7] For the state groupings in the regions, it helps with the statistical sizes for the tests that are going to be done. It also helps to identify some similarities and differences between the geographical regions. For each state and region, counts are divided by the population of that area to get the per capita value. The population data in each state is from the US Census 2017 State Estimates. [2] To determine the population for the regions, the state populations for each region are added together to get the total region's population. [7] This will be the same process to find the positive cases, hospitalized patients, and new deaths for each region.

Table 2.0-1, States Contained in each Region, contains a row for each Region and the states contained within each of the four regions. The District of Columbia is considered a separate state with respect to the data being analyzed and is contained in the South Region. [7]

Table 2.0-1, States Contained in each Region

Region (#States)	States in Region
Midwest (12)	Indiana, Illinois, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota and South Dakota.
Northeast (9)	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York, and Pennsylvania.
South (17)	Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma, and Texas.
West (13)	Arizona, Colorado, Idaho, New Mexico, Montana, Utah, Nevada, Wyoming, Alaska, California, Hawaii, Oregon, and Washington.

3.0 Analysis

The COVID data being used is analyzed using Functional Data Analysis (FDA). [13, 3] The COVID data will be analyzed using the RStudio program. [4, 8, 14] FDA is the process of looking at different smoothed curves and analyzing the curves. With the different curves, there are functions of data over time. The domain for most of the data is over time but the data can have different domains if the data can be plotted on a graph to analyze. When the data is over time, there is a point for each day for the length of time. Once there is a point for each time interval all the points are connected to create a functional data curve.

Functional data uses basis functions which are a set of functional building blocks. The constant basis system is the single basis function 1, which is often needed alone. Splines and Fourier series are with two or more basis systems. Non-periodic functions are functions that do not remain the same throughout. Periodic functions are functions that follow the same pattern every time and remain unchanged. Periodic functions can use either the spline basis system or the Fourier basis functions to approximate the function. Non-periodic functions use spline basis systems or B-spline systems to explain the data. The COVID analysis uses periodic functions with the Fourier basis functions to explain the data and look at it further.

For this data, there are different and multiple replications. The data could have fifty-one different replications for each of the states (plus the District of Columbia) or four replications for the regions. When the data has different replications sometimes the lines are very rough or jagged and not a smoothed line over time. The next part of the FDA is to smooth the functional data line over time.

The generalized cross validation score (GCV) helps to find the best lambda for smoothing the data, which is determined and explained later.

Once the functional data lines are smoothed over time, the functional data can be analyzed for features of the curves. The replications of the curves over time can be analyzed to find similarities and to create predictions of the data. It can find the mean, variances and covariances of the data and compare that data with the other replications. The functional principal component analysis (FPCA) of the data can be found and analyzed between the replications.

The COVID-19 data being used is over time with 4 replications for each region or 51 replications for each state (plus the District of Columbia). The data that is examined is the:

- number of positive cases over time,
- number of hospitalized patients over time,
- number of deaths over time.

These cases of data will be examined by each of the 4 regions and also by each of the 50 states and the District of Columbia.

3.1 New COVID Cases Analysis

Section 3.1 New COVID Cases Analysis, looks at the new COVID cases for each state and region by day. Here, the optimal smoothing for the data is found and applied. Future sections look at the contour and correlation coefficient plots to determine during which months the most positive COVID cases are identified per region and state. This section, as well as future sections, will also look at the following:

- eigenvalues,
- cumulative percentage,
- functional principal component functions,
- beta graphs,
- regression coefficients,
- functional F-Test.

Analyzing the state and region data utilizing these different analysis methods will help compare the new cases by state and region then come to a valuable conclusion about the data.

For this data, the following are the null and alternative hypotheses:

- Null Hypothesis: H_0: There is no difference in the regions for new COVID cases
- Alternative Hypothesis: H_a: There is a difference in the regions for new COVID cases.

3.1.1 New Cases by Region and State

The first data set analyzed is the New COVID-19 Cases by each region and state. In Figure 3.1.1-1, New COVID Cases per Region by Day, displays for each region the number of new COVID cases for that day by the population.

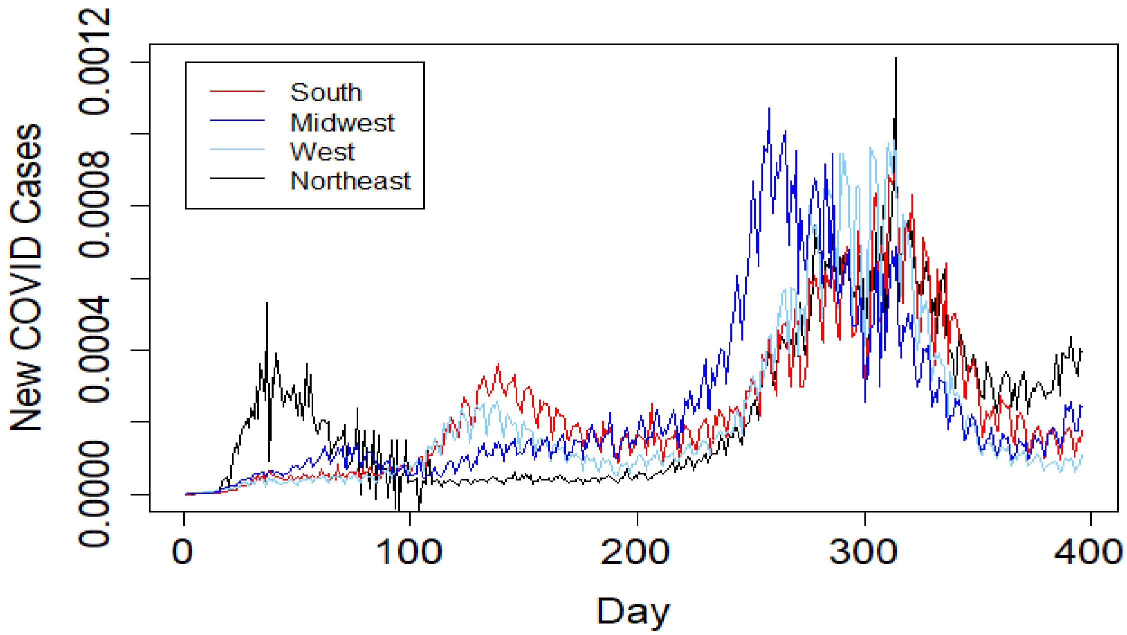

Figure 3.1.1-1, New COVID Cases per Region by Day

Figure 3.1.1-1 shows that the number of new cases starts low in all four regions, then in the Northeast region new cases spike up first. Looking at day 100 we see that the South and the West spike up and then go back down at day 200. Around day 250 all regions spike up and then go back down to be about the same constant number at day 400. The Northeast region has the highest new COVID Cases around day 310 comparing all the regions.

Figure 3.1.1-2, New COVID Cases per State by Day, displays the number of new positive cases per day by each region by the color and then the different lines are the states.

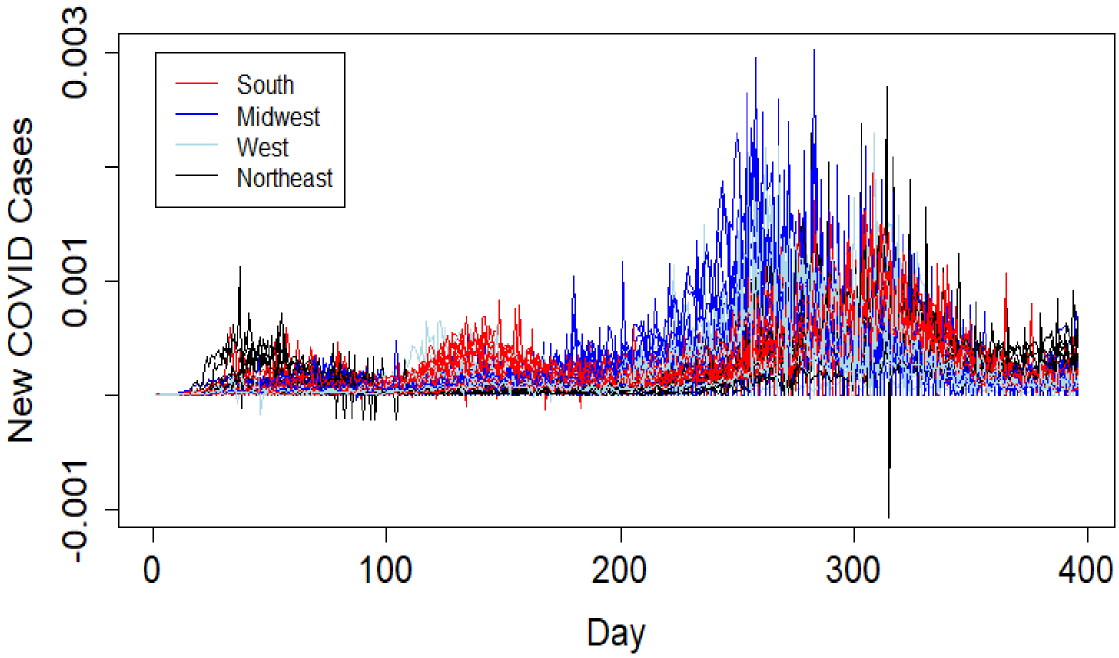

Figure 3.1.1-2, New COVID Cases per State by Day

When looking at Figure 3.1.1-2, all the lines on the graph seem to be following the same trends. There are some states that are higher than others, but the new COVID Cases per state all seem to follow the same trend. In the graph, there is one state that stands out because the state has negative COVID cases. There are a few states that have negative cases for certain days but this one has the most and is an outlier, which is discussed later. The negative cases could have come from some errors or delays in the data. It also could have come from the states revising their case data by removing duplicates or non-confirmed cases that were reported already.

Figure 3.1.1-3, New COVID Cases per Region by State by Day, displays the four regions and for each region, the states that are noticeable (have high or low COVID Cases) in that region.

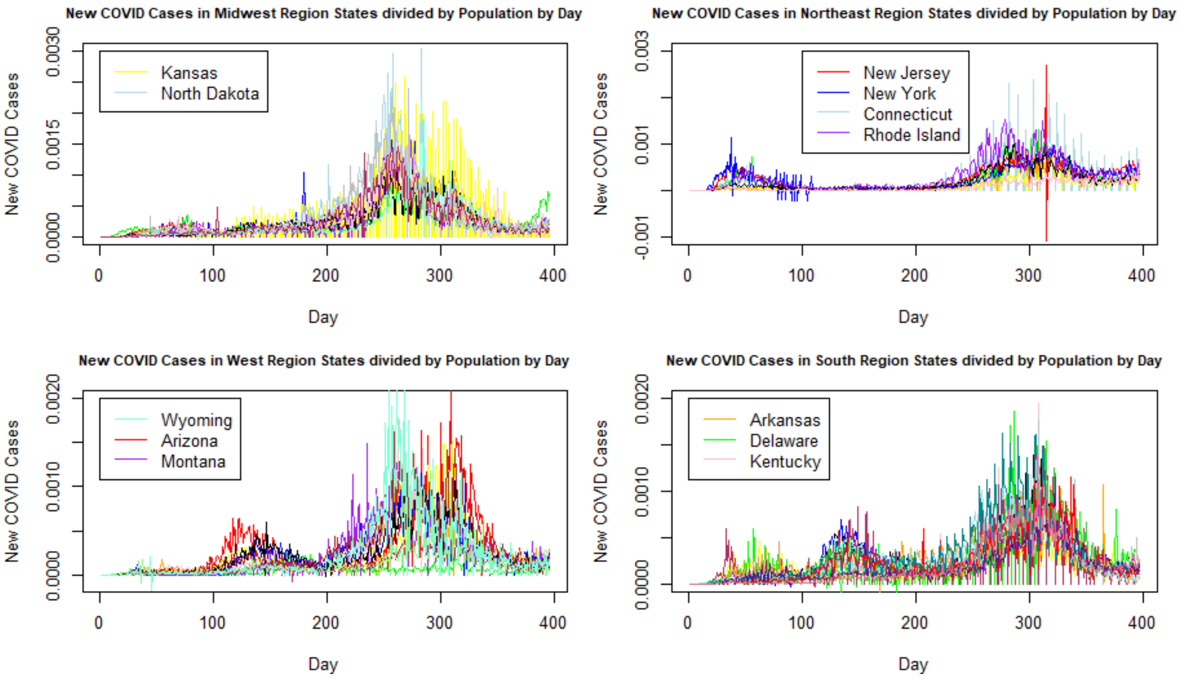

Figure 3.1.1-3, New COVID Cases per Region by State by Day

Looking at the Midwest Region States first, Kansas and North Dakota have the highest number of new COVID Cases. In the Northeast Region, New Jersey, New York, Connecticut, and Rhode Island are the states that have some spikes either positive or negative. New Jersey stands out the most with a negative spike downwards. In the West Region the states that stand out the most are Wyoming, Arizona, and Montana. There are some spots where these states stand out but most of the states follow along in the same pattern. In the South region Arkansas, Delaware, and Kentucky are the states that stand out the most. There are some spikes in the data for some of the states. Most of them in this region also follow along the same pattern.

3.1.2 Smoothed Line New Cases

To smooth the data, the best value of the smoothing parameter (lambda) is first determined. When smoothing the data, the same lambda is used for all 3 sets of data:

- New COVID cases,
- New hospitalized patients, and
- New deaths.

Using the same lambda for all three data sets will help keep the data accurate and easier to understand. Figure 3.1.2-1, Generalized Cross Validation of Lambda, displays the generalized cross validation (GCV) graphs for the new COVID cases, hospitalized patients, and new deaths.

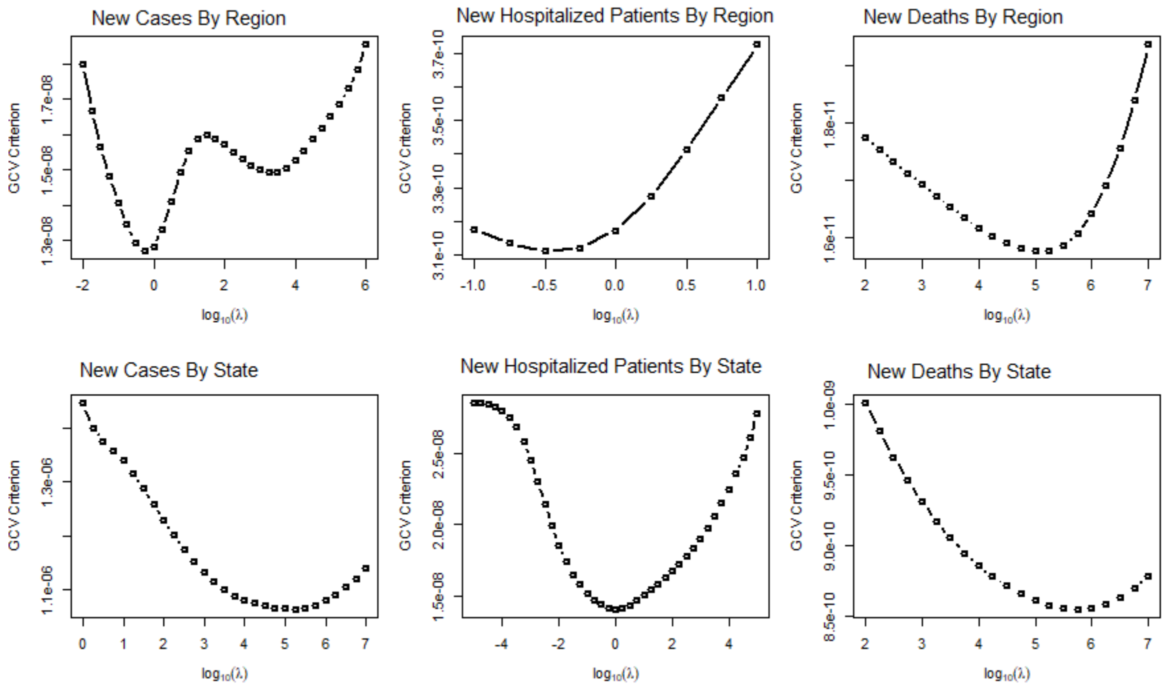

Figure 3.1.2-1, Generalized Cross Validation for Lambda

Figure 3.1.2-1 shows what the best lambda is for each set of data, new COVID cases, new hospitalized patients, and new deaths. When the graph is at its lowest point is what the best lambda is for that data set. Depending on which data is evaluated, the best lambda is between 0 and 6. For this data a lambda value of 4 is selected, which is in the middle and helps smooth out all the data sets. Figure 3.1.2-2, Smoothed Line of New COVID Cases per Region by Day, illustrates why a lambda of 4 is selected for the data.

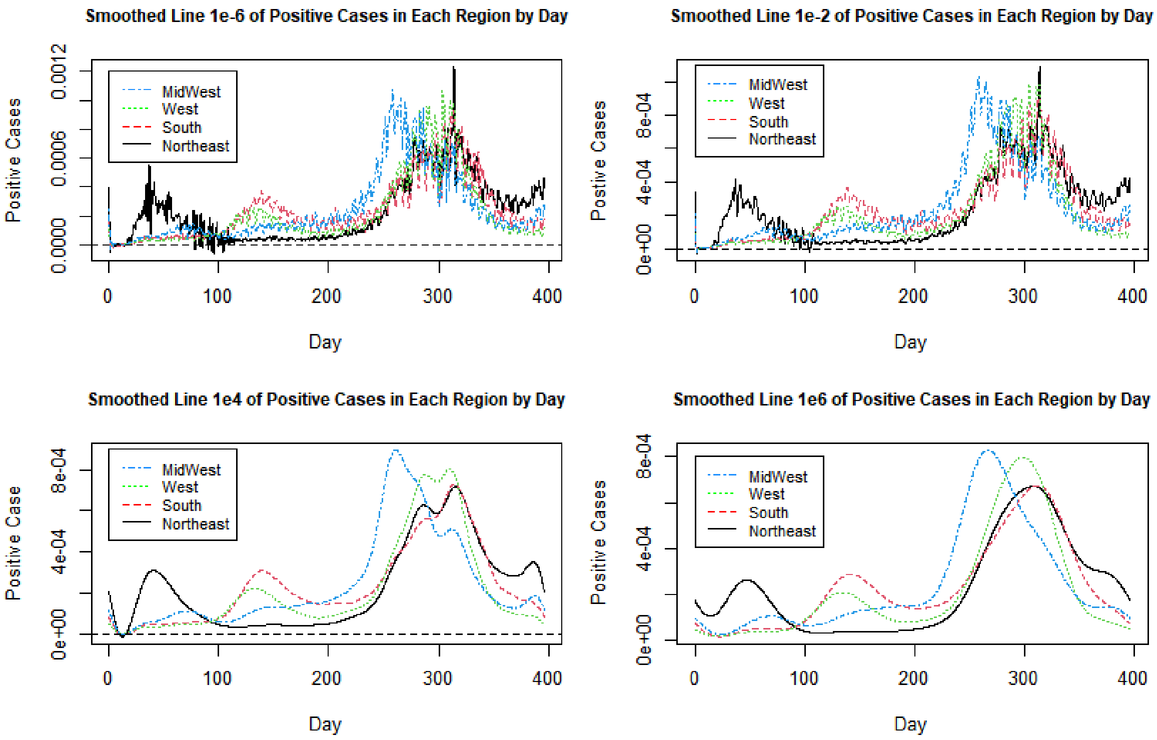

Figure 3.1.2-2, Smoothed Line of New COVID Cases per Region by Day

In Figure 3.1.2-2 there are four different smoothed line graphs to evaluate and determine which one is the best fit for this data. The first smoothed line graph is 10^{-6}. Here the graph shows lines still spike up and down with the data. The line needs to be more smoothed to be able to see the trend lines better. The second graph smooths the line out more to 10^{-2}, but still has some up and down spikes and could look better. The next graph is the smoothed line 10^{4}, which in Figure3.1.2-1 was in the middle between the best lambdas selected. With the smoothed line of 10^{4}, this graph looks a lot better with the data and lines smoothed out more. The data is smoothed out one more time using a lambda of 1e6. For this lambda value, the graph is too smooth. The data needs to have some curves, to be able to see what is going on and to be able to analyze the data better. When looking at Figure 3.1.2-1 and Figure 3.1.2-2 the best lambda that will be utilized in this data analysis thesis is 1e4.

Figure 3.1.2-3, Smoothed Line 10^{4} of New COVID Cases per Region by Day, illustrates the smoothed lines of the new COVID Cases per Region divided by the region's population by each day. Figure 3.1.2-3 is the Smoothed Line graph in the prior Figure 3.1.2-2 but enlarged to show better details.

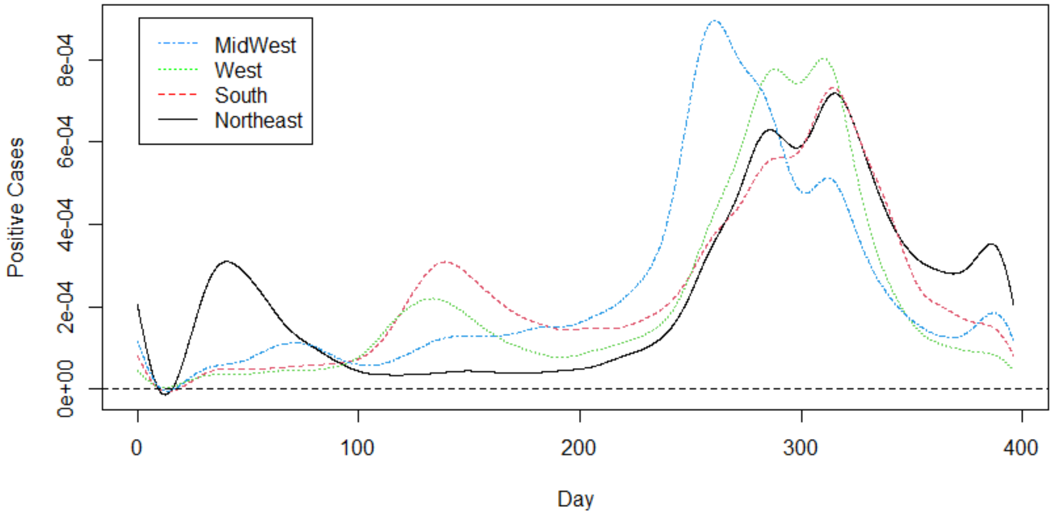

Figure 3.1.2-3, Smoothed Line of New COVID Cases per Region by Day

Figure 3.1.2-3 with the smoothed line of the new COVID cases per region divided by the population in that region, the Northeast area spikes up right away and then goes back down. Then around day 250 they all spike up with the Midwest area having the highest cases out of the other regions.

Figure 3.1.2-4, Smoothed Line of New COVID Cases per State, the new cases by states divided by that state's populations is being displayed.

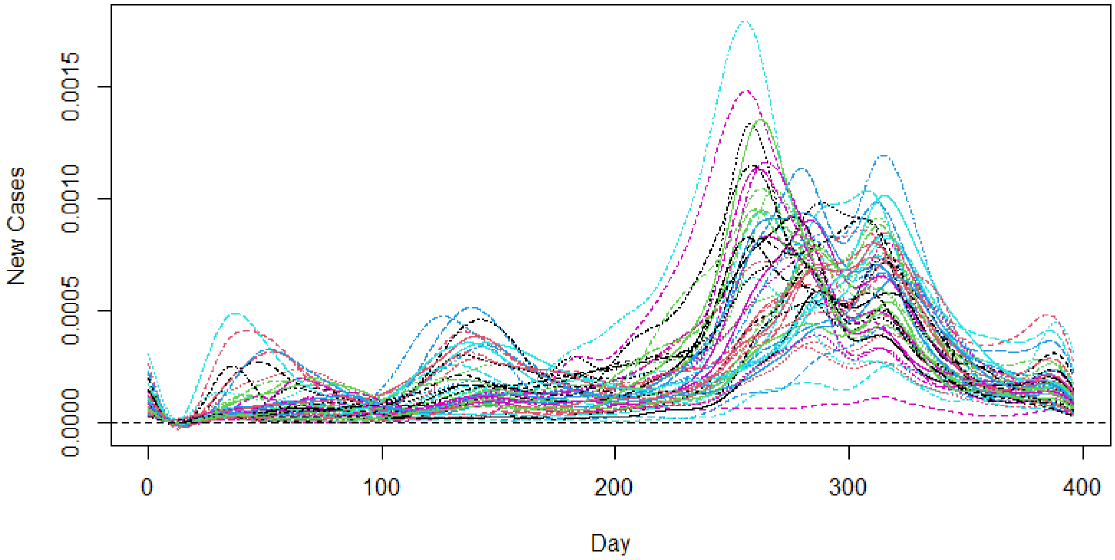

Figure 3.1.2-4, Smoothed Line of New COVID Cases per State

In Figure 3.1.2-4, which is the smoothed line of the new COVID cases per state by day, one state stands out more compared to the other states around day 250, which is discussed later. The new cases follow along the same paths and there aren't many states that stand out from the others.

3.1.3 Smoothed Line Graph of New Cases

This section looks at the new COVID cases over time and the average over time for all the data. Figure 3.1.3-1, Smoothed Line of New COVID Cases with the mean Line per Region by Day, shows the number of positive cases in that region for that day over time.

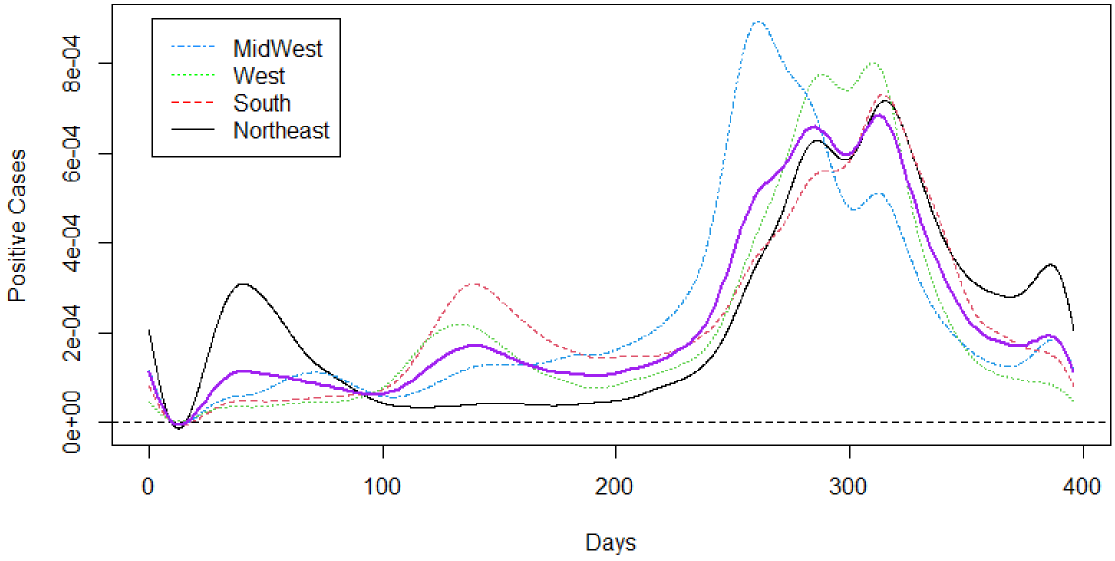

Figure 3.1.3-1, Smoothed Line of New COVID Cases with the Mean Line per Region by Day

The purple line shows the mean positive cases for all the regions over time for that day. The average line looks to be in the middle of the four regions, following the same trend that the other data is. Around day 50 the Northeast region pulls the average up to make it higher than the other regions. Around day 250 the Midwest is up a lot higher than the other regions.

Figure 3.1.3-2, Smoothed Line of New COVID Cases with the Mean Line per State by Day, is the smoothed lines of the new COVID. On this graph, the mean line of the data is plotted in a solid bold purple line as well as the COVID cases by states lines. The mean line of the data will help us analyze where the average of the data is and see which states stand out from the average.

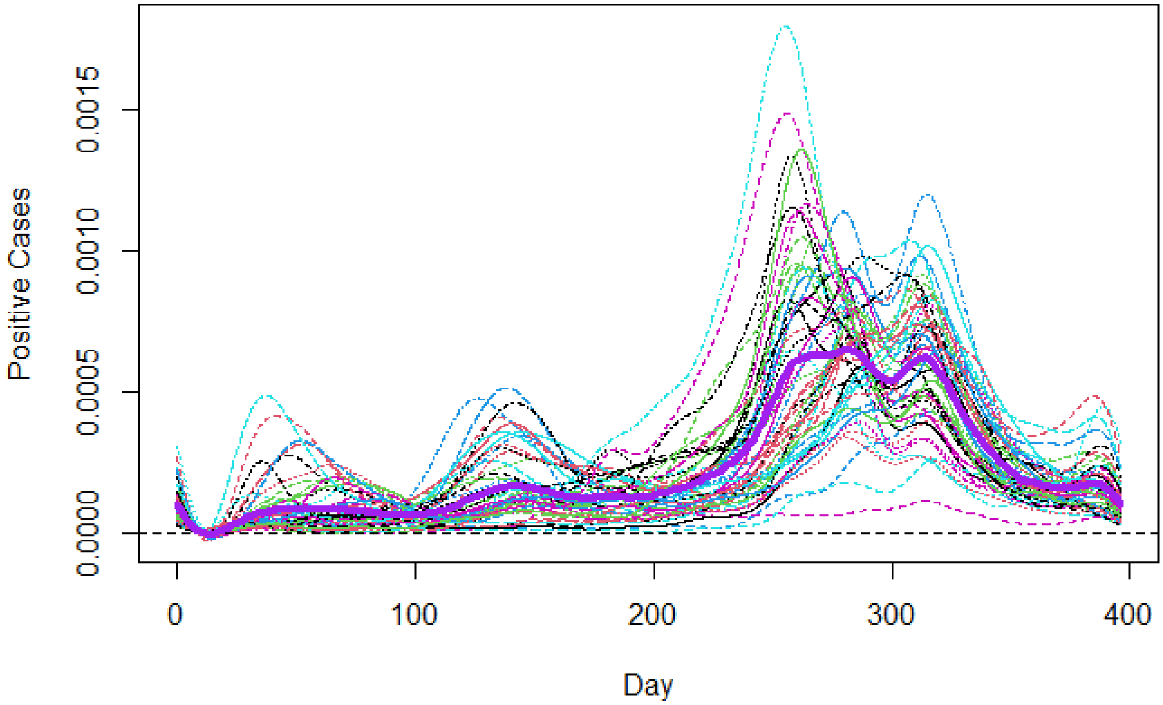

Figure 3.1.3-2, Smoothed Line of New COVID Cases with the Mean Line per State by Day

Figure 3.1.3-2 shows that around days 50, 150, 250, and 320 there are spikes in the data. For those days the mean line goes up but not as high as some of the data does. There are some states that pull the mean line up, compared to the other states that have a lower new COVID case count.

The data can also be displayed as smoothed lines minus the mean line. This helps to see when each region is above or below average over time.

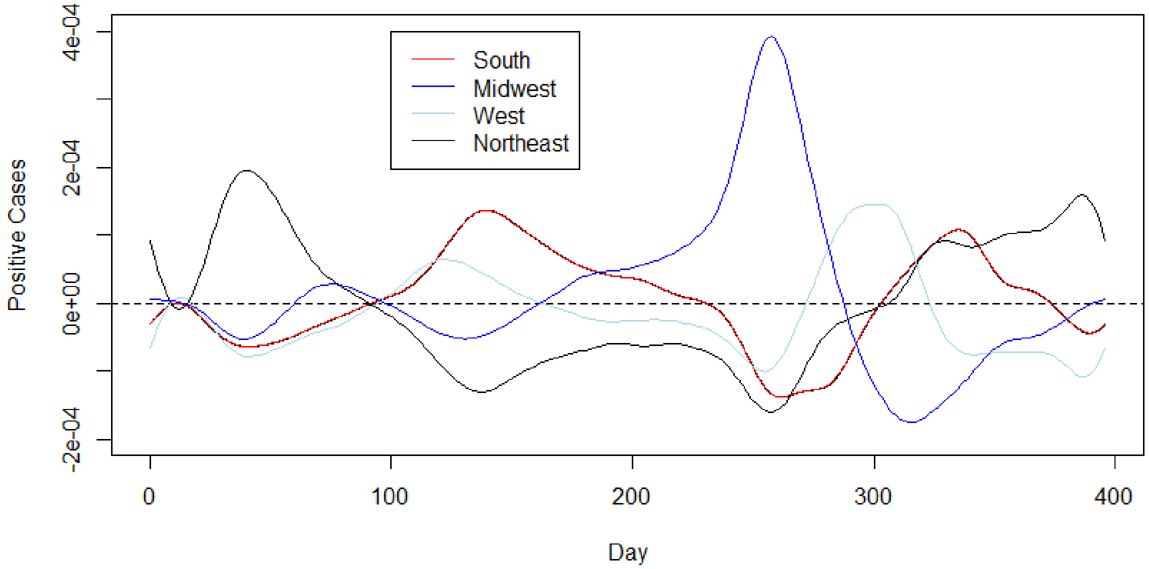

Figure 3.1.3-3, Smoothed Line of New COVID Cases minus the Mean per Region by Day

Figure 3.1.3-3 shows that the Northeast area is again the highest in the beginning like it shows in Figure 3.1.3-1. The Northeast region is a lot higher at first, then drops down below the rest and comes back up at the end. The South stays constant going up and down over time but nothing to concerning. The Midwest stands out the most around day 250 that spikes up higher than the rest of the regions.

The last Figure 3.1.3-4, Smoothed Line of New COVID Cases minus the Mean for States by Day, shows the number of cases in each region by the states minus the averages.

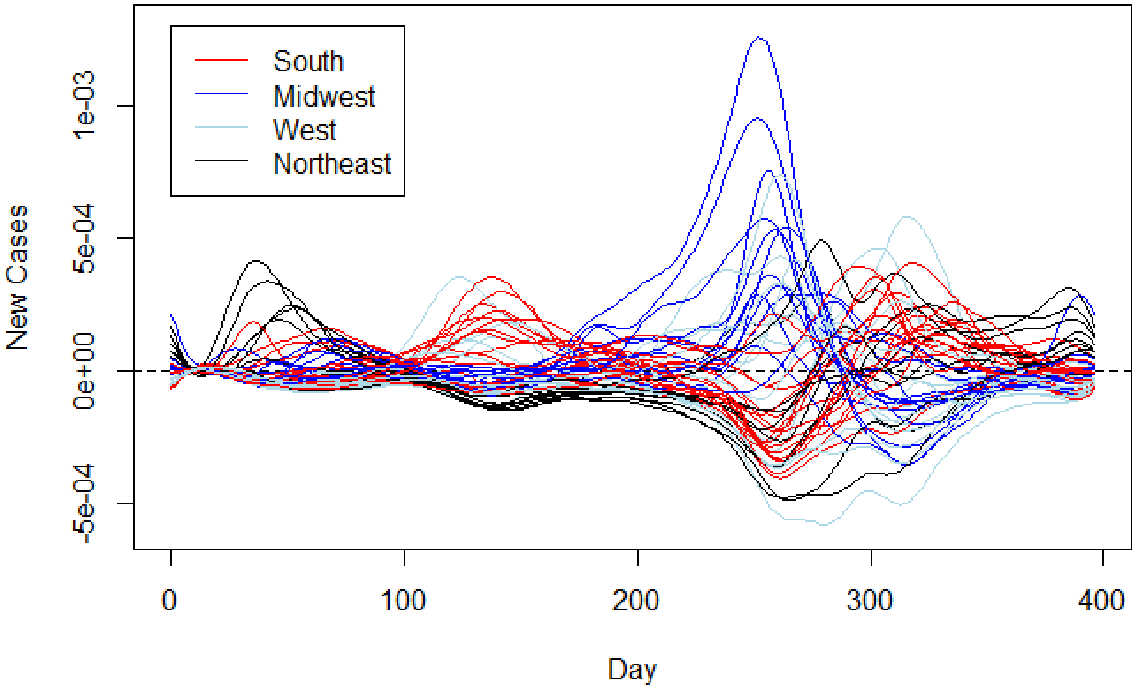

Figure 3.1.3-4, Smoothed Line of New COVID Cases minus the Mean for States by Day

Looking at the last Figure 3.1.3-4 in the beginning there are a few Northeast states that are above the average that stand out. Around day 150 it switches, and the South states stand out above the average and the other states. Then around day 250 the Midwest states are the highest and really stand out above the rest. From there as time goes on the states flatten out to be about the same around the 0 and the average.

3.1.4 Variance-Covariance of New Cases

In Figure 3.1.4-1 shows the contour plot of the variance-covariance of the new COVID cases per region by day.

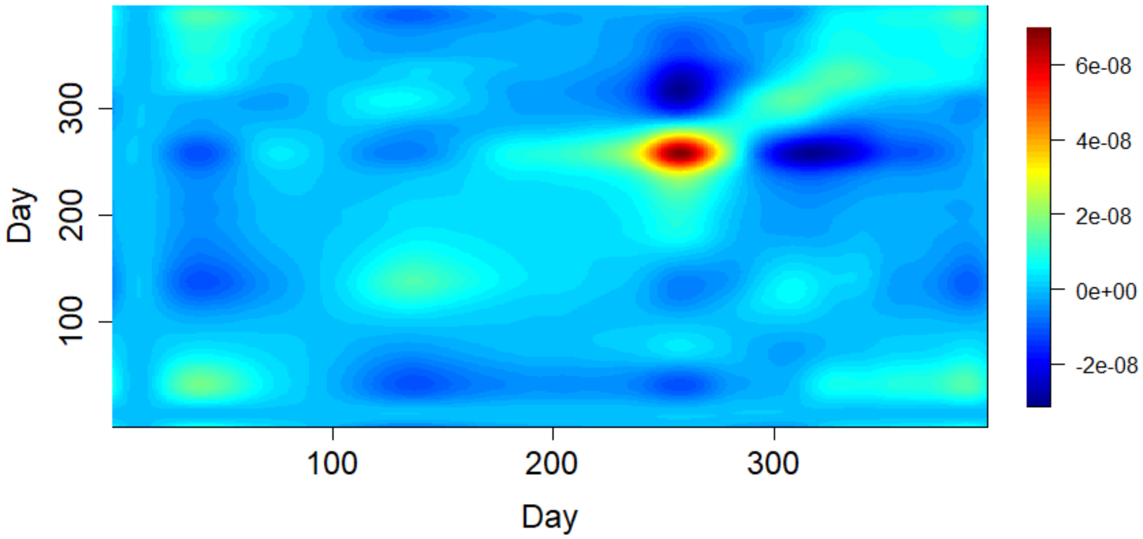

Figure 3.1.4-1, Contour Plot of New COVID Cases per Region by Day

The diagonal of the plot is also the variance of the data. On the diagonal of the plot, for each day the variance counts are a mirror image of each other. The dark red area on the plot is the highest COVID count which is 6e-08, and is along the diagonal which is part of the variance.

Figure 3.1.4-2, Contour Plot of New COVID Cases for States by Day, displays a contour plot of the variance-covariance of the new COVID cases for states by day.

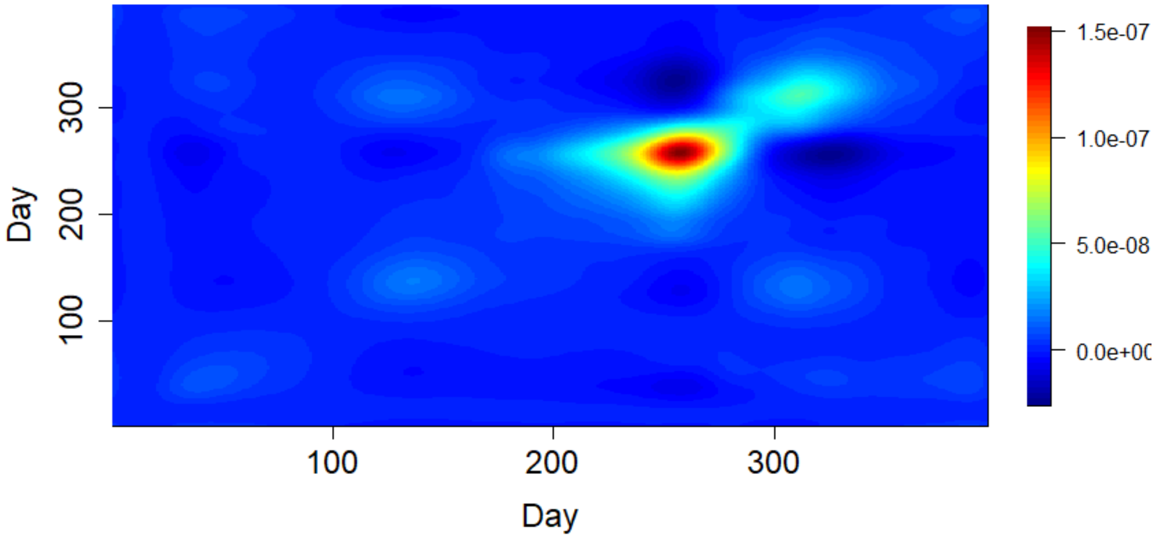

Figure 3.1.4-2, Contour Plot of New COVID Cases for States by Day

This plot shows the covariance surface of the New COVID cases for states by each day. The diagonal of the plot is the variance of the data. Across the diagonal of the plot on each site they mimic each other. The dark red of the plot is the highest count which is 1.5e-07. This part is along the diagonal which is part of the variance. The next section examines the correlation coefficient plots. These are easier to interpret than covariances and to tell more from the data.

3.1.5 Correlation Coefficient Plot New Cases

This section investigates the correlation coefficient plots of the new COVID cases over time. The X and the Y axis are both days to compare the time of year and the number of COVID cases there have been. Figure 3.1.5-1, Correlation Coefficient Plot of New COVID Cases per Region by Day, displays the correlation coefficient plot of the new COVID cases per region by day.

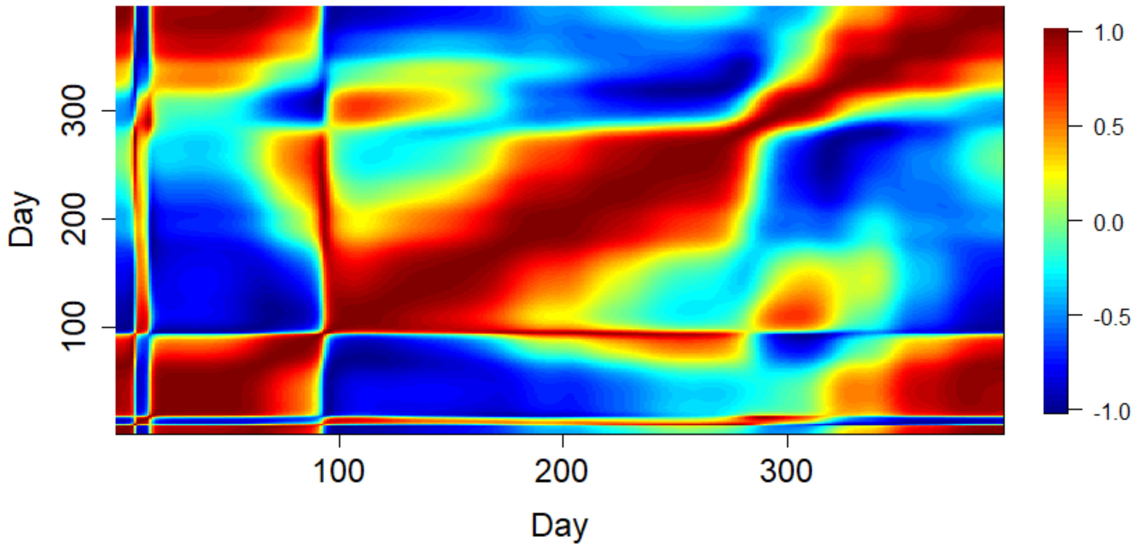

Figure 3.1.5-1, Correlation Coefficient Plot of New COVID Cases per Region by Day

Figure 3.1.5-1 is easier to read and conveys more than the variance-covariance plot that is in Figure 3.1.4-1, Contour Plot of New COVID Cases per Region by Day, above. Figure 3.1.5-1, along the diagonal, is equal to the value 1 which is the maximum correlation. The highest value is along the diagonal when it is 1, the lowest correlation is the darker blue (-1.0) which is the correlation between June to July (x-axis) and April to June (y-axis) in the region areas. The lower correlation represents (dark blue) inverse relations between the variables. Showing that during that time there is less likelihood of having new COVID cases in that region at that time. As opposed to the dark red area that shows that during that time there is more likelihood of having new COVID cases in the regions at that time.

Figure 3.1.5-2, Correlation Coefficient Plot of New COVID Cases for States by Day, is the correlation coefficient plot of the new COVID cases for states by day. Along the diagonal is again, the highest correlation value of 1. The smallest correlation is in the dark blue again of -0.5, which is the correlation from March to April in the state's areas.

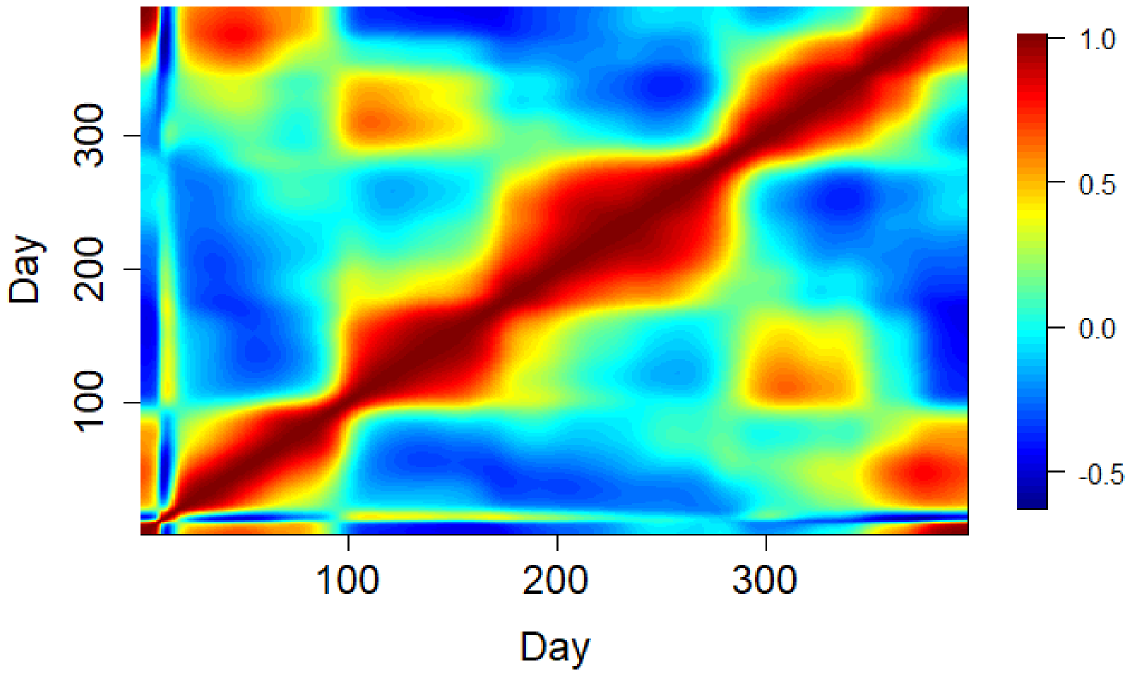

Figure 3.1.5-2, Correlation Coefficient Plot of New COVID Cases for States by Day

3.1.6 Eigenvalue Plot New Cases

This section analyzes the data using Functional Principal Component Analysis (FPCA) and the eigenvalues of the new COVID cases. All the data sets use 4 Functional Principal Components (FPC) which is discussed more in section 3.1.8. Figure 3.1.6-1, Eigenvalue Plot of New COVID Cases per Region by Day, looks at each eigenvalue for each FPCA.

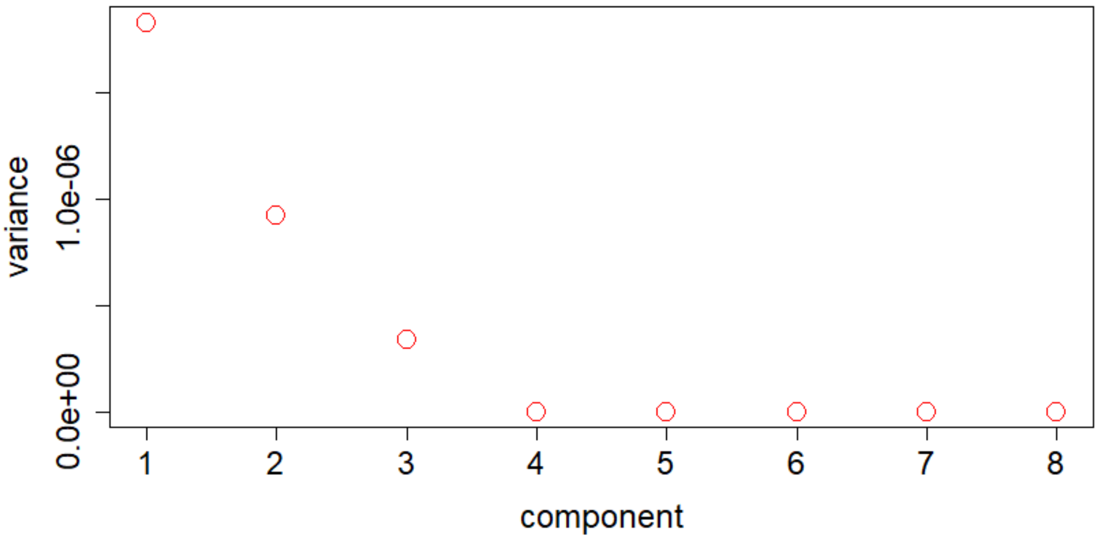

Figure 3.1.6-1, Eigenvalue Plot of New COVID Cases per Region by Day

Each eigenvalue is equal to the variance of the FPC score. As the eigenvalue component increases the variance value decreases and gets closer and closer to 0. The first eigenvalue is large and then the next one drops in value very fast. Figure 3.1.6-2, Eigenvalue Plot of New COVID Cases for States by Day, look at the same values but for the new COVID cases for states by day.

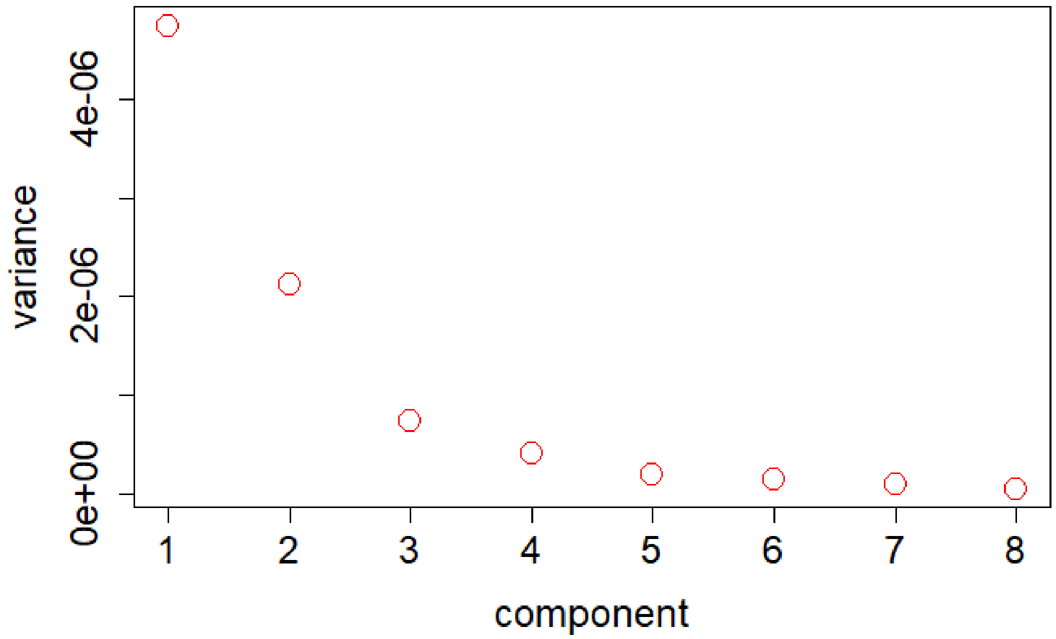

Figure 3.1.6-2, Eigenvalue Plot of New COVID Cases for States by Day

The values and eigenvalues for the new COVID cases for states by day are bigger values than the ones for new COVID cases per region. The eigenvalues for Figure 3.1.6-2 drop a little slower than in Figure 3.1.6-1. The eigenvalues don't seem to get to around 0 until the 5th eigenvalue.

3.1.7 Cumulative Percentage Plot New Cases

This section looks at the cumulative percentage explained for the total variation plots of the new COVID cases. These plots are the sum of the ten eigenvalues then divided by the sum of all the eigenvalues. This shows what number of FPC's (x-axis) are needed to show what percentage of the data (y-axis).

Figure 3.1.7-1, Cumulative Percentage Plot of New COVID Cases per Region by Day, displays the cumulative percentage of the new COVID cases per region by day.

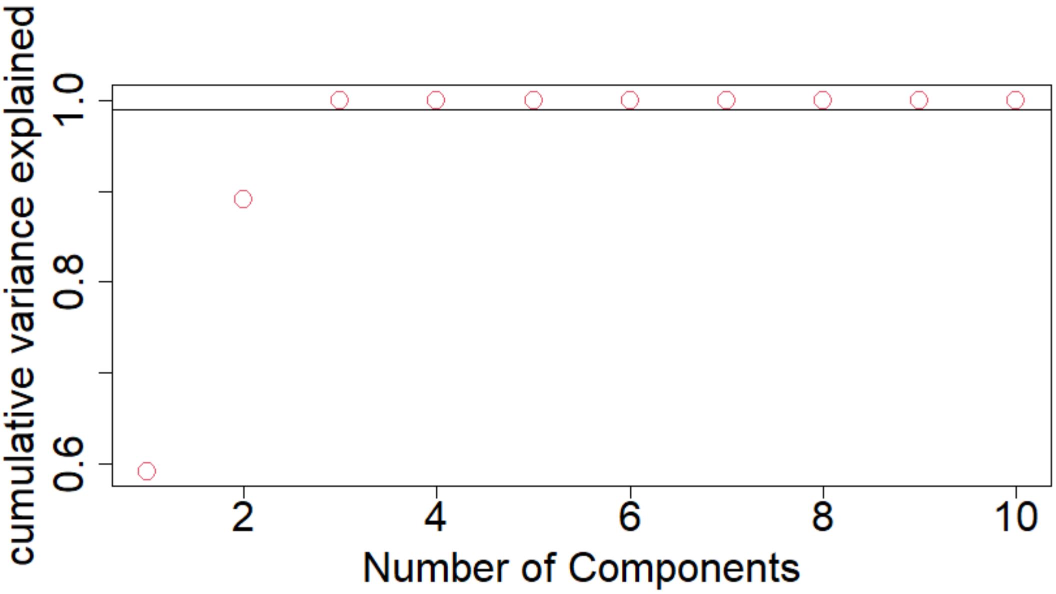

Figure 3.1.7-1, Cumulative Percentage Plot of New COVID Cases per Region by Day

Figure 3.1.7-1 shows that the first 3 FPCS explain about 99% of the total variation of the data. Once the graph reaches the 4th and above FPCS the graph shows that the value is above the 99% line and remains about the same value. With this data the first 3 FPCs are needed to explain most of the data.

Figure 3.1.7-2, Cumulative Percentage Plot of New COVID Cases for States by Day, displays cumulative percentage of the new COVID cases for states by day.

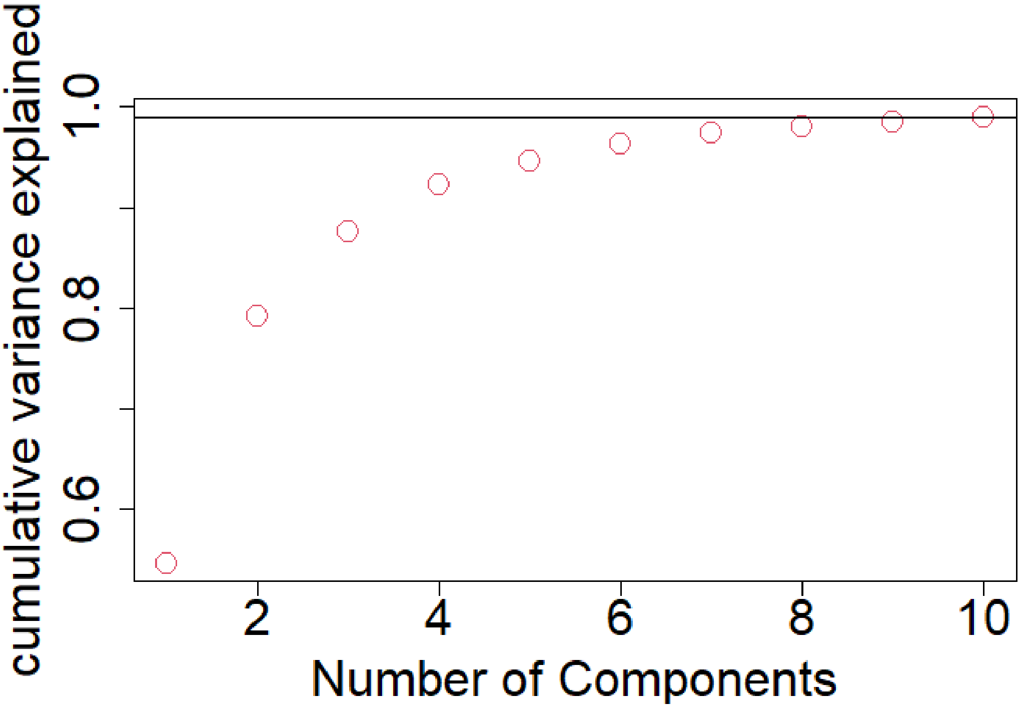

Figure 3.1.7-2, Cumulative Percentage Plot of New COVID Cases for States by Day

Figure 3.1.7-2 displays that more FPCs are needed to explain the data than in Figure 3.1.7-1. For the new COVID cases for states by day, the number of components does not go above the 1 line until about 10 components. This data would need about 4 or 5 FPCs to explain about 99% of the data.

3.1.8 Functional Principal Component Functions Graph of New Cases

Section 3.1.7 presented cumulative percentages and how many FPCs are needed to explain the data. Looking at Figures 3.1.7-1 it shows that 3 are needed and 3.1.7-2 shows that 4 or 5 are needed. For sections 3.1.8, 3.2.8 and 3.3.8, 4 FPCs are used to stay consistent and to look further into the data.

Figure 3.1.8-1, Functional Principal Component Functions of New COVID Cases per Region by Day, looks deeper into explaining the data.

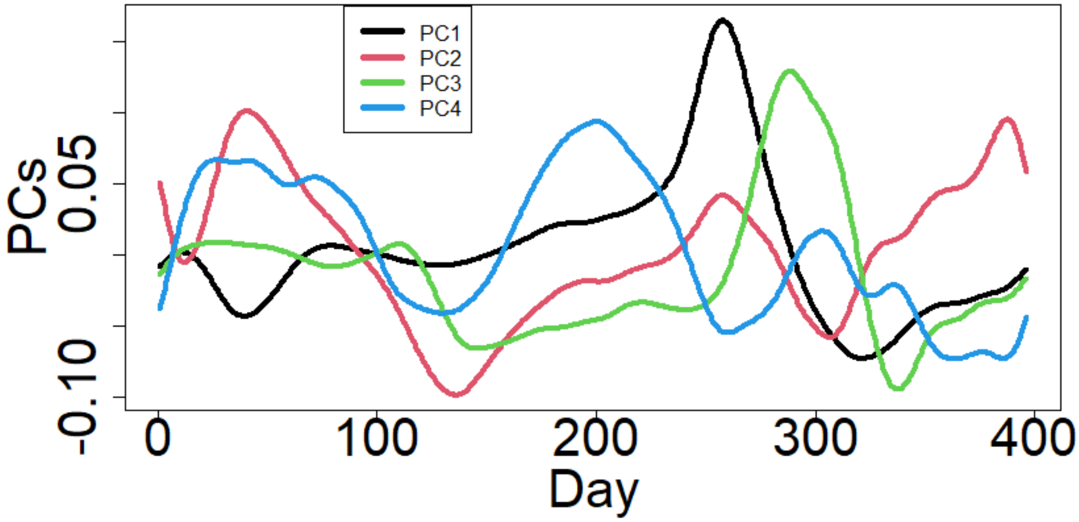

Figure 3.1.8-1, Functional Principal Component Functions of New COVID Cases per Region by Day

In Figure 3.1.8-1, the first principal component (PC1), which is the black line, is the weighted average of the new COVID cases. When PC1 begins at day 0, it has a lower number which matches up with the new COVID cases per region day at that time. As time goes on PC1 spikes up around day 260, and then drops down around day 320, which also matches up with the new COVID cases per region over time. The Midwest Region new cases follow along the same black line curve as PC1. The second principal component (PC2), which is the red line, shows the change in the COVID cases for the different months or seasons. The PC2 is a positive value during the spring and then drops and is negative most of the summer, fall, and winter months, then goes back up for the spring. PC2 matches up the best with the Northeast region, spiking up in the beginning then falling and then spiking up again at the end. PC3 matches up best with the west region, staying lower in the beginning and then spiking up around the end. After PC3 spikes up it goes back down lower in the end. The last PC4 then fits best with the south region.

Figure 3.1.8-2, Functional Principal Component Functions of New COVID Cases for States by Day, displays the four Principal Components that best represents/explains the data.

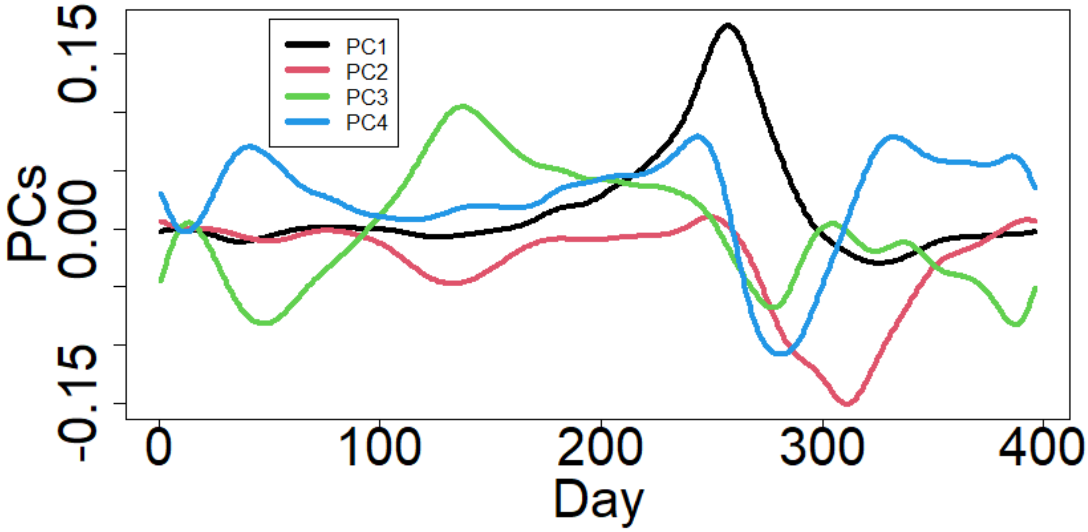

Figure 3.1.8-2, Functional Principal Component Functions of New COVID Cases for States by Day

Figure 3.1.8-2 is similar to Figure 3.1.8-1 (which represents Region Data), where Figure 3.1.8-2 represents data for each state and not just the regions. This figure needs the states that match up best with each principal component. Referencing Figure 3.1.1-3, New COVID Cases per Region by State by Day, some of the states that stand out match up with some of the Principal Components (PC). New York matches up best with PC4, Kansas with PC1, North Dakota with PC1, Montana with PC3, and New Jersey with PC2. All of the states match up with one of the PC's but those examples just show some of the states that stood out from the others.

3.1.9 FPC1 vs FPC2 Graphs of New Cases

This section investigates the FPC1 on the x-axis and the FPC2 on the y-axis. This shows what states and regions are grouped together and are similar and which ones stand out from the others. Figure 3.1.9-1, FPC1 vs FPC2 Scores of the New COVID Cases per Region by Day, displays FPC1 (x-axis) against FPC2 (y-axis).

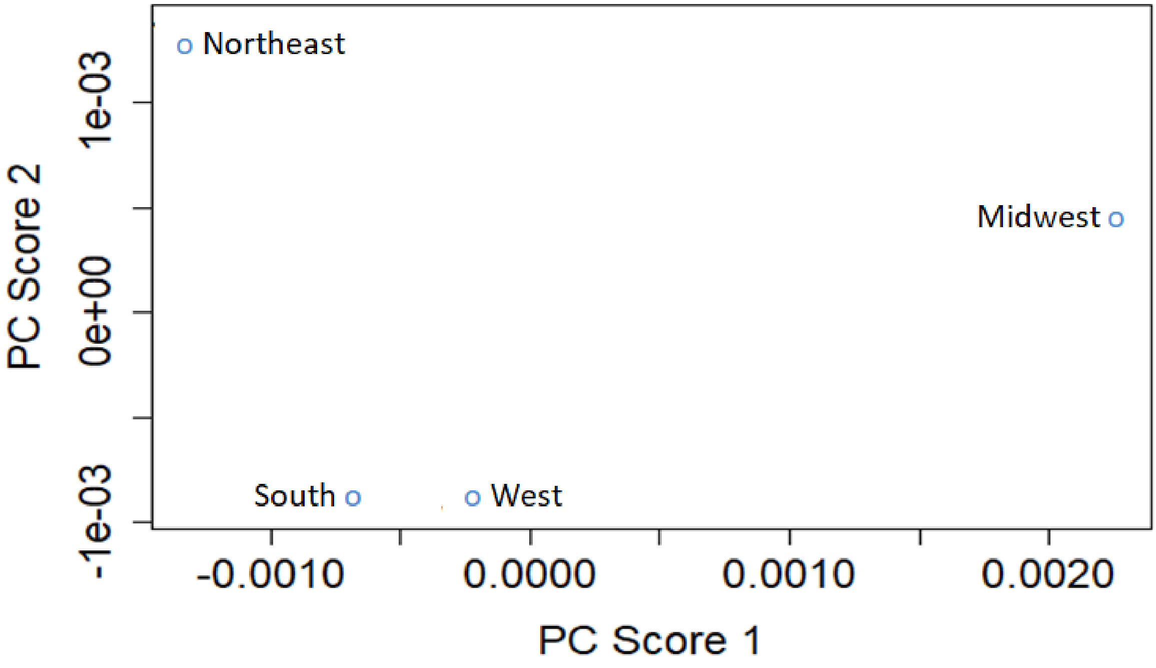

Figure 3.1.9-1, FPC1 vs FPC2 Scores of New COVID Cases per Region by Day

Looking at Figure 3.1.9-1, one can see that the South and the West regions are grouped together. With those regions grouped together, it shows that those two regions are similar with new COVID cases. The Northeast and the Midwest are not grouped with any other region, showing that these regions aren’t like the other regions in the new COVID cases. Figure 3.1.9-2 looks at the same type of graph but with the different states, showing which states are similar in the new COVID cases. It also shows which states are outliers and aren’t similar to the other states.

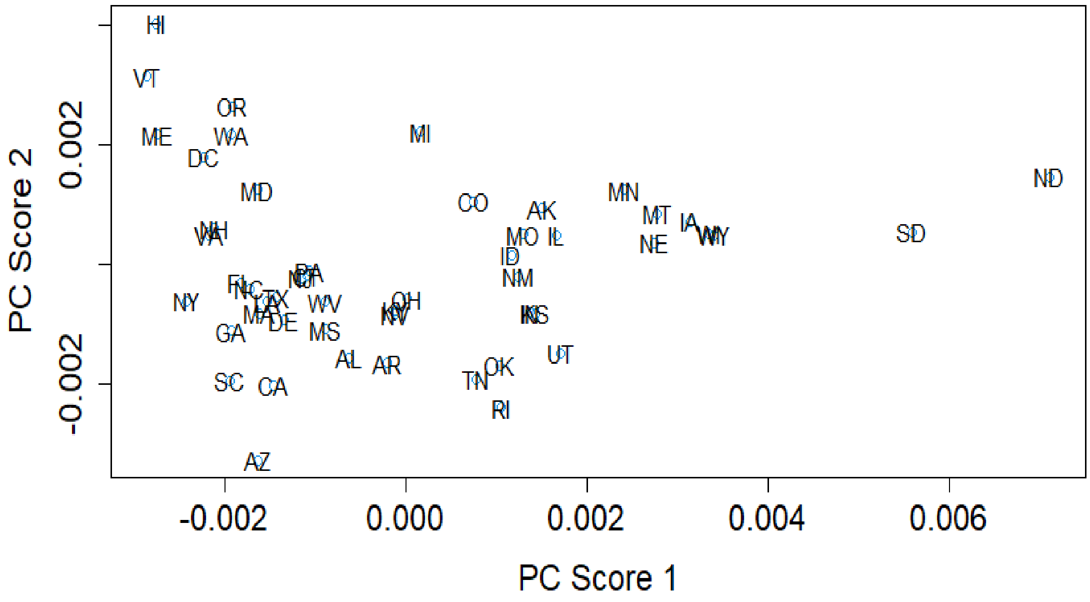

Figure 3.1.9-2, FPC1 vs FPC2 Scores of New COVID Cases for States by Day

In Figure 3.1.9-2, most of the states are grouped together but there are a few that stand out from the others. North Dakota and South Dakota are off to the right from the other states, and this shows that those states didn’t follow along with the other states in the new COVID cases. There are a few states that are higher and lower than the other states like Hawaii and Arizona, but those states don’t stand out as much as North Dakota and South Dakota do.

3.1.10 Regression Coefficients

Figure 3.1.10-1, New COVID Cases Mean, Difference and Predicted by Region by Day, displays the regression coefficients and the differences between the mean and each region new COVID cases.

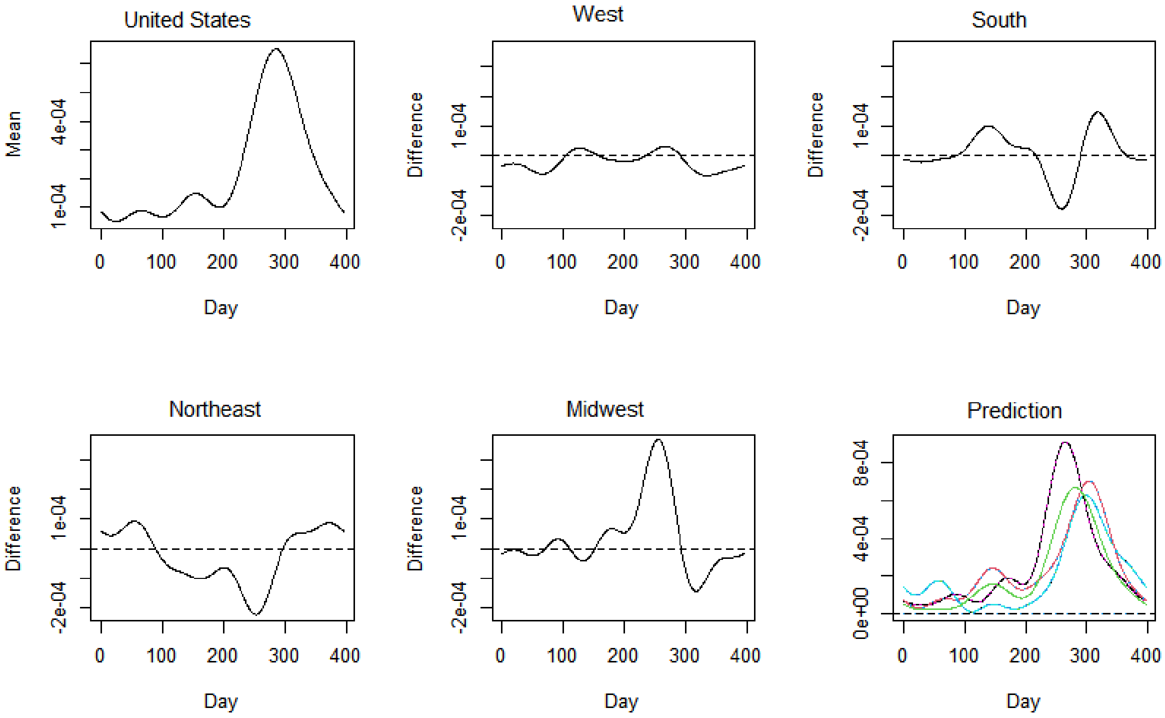

Figure 3.1.10-1, New COVID Cases Mean, Difference and Predicted by Region by Day

The United States graph in Figure 3.1.10-1 shows the mean of all the new COVID cases over the time frame. The next four graphs show the difference in that area of the new COVID cases and the mean from the United States. With the four regions what they are in the positive values it shows when they have the highest impacts. When the values are in the positive it shows that they had more positive

COVID cases in that region and what the mean is for that time. The graph shows the predictions or the positive COVID cases for that region.

3.1.11 Functional F-Test (ANOVA)

This last section looks at the Functional F-Test for the new COVID cases over time. As was noted earlier, for this data, consider the hypothesis:

- Null Hypothesis: H_0: There is no difference in the regions for new COVID cases
- Alternative Hypothesis: H_a: There is a difference in the regions for new COVID cases.

Figure 3.1.11-1, New COVID Cases for Functional F-Test Statistic by Day, shows if there is a difference between the regions with the new COVID cases.

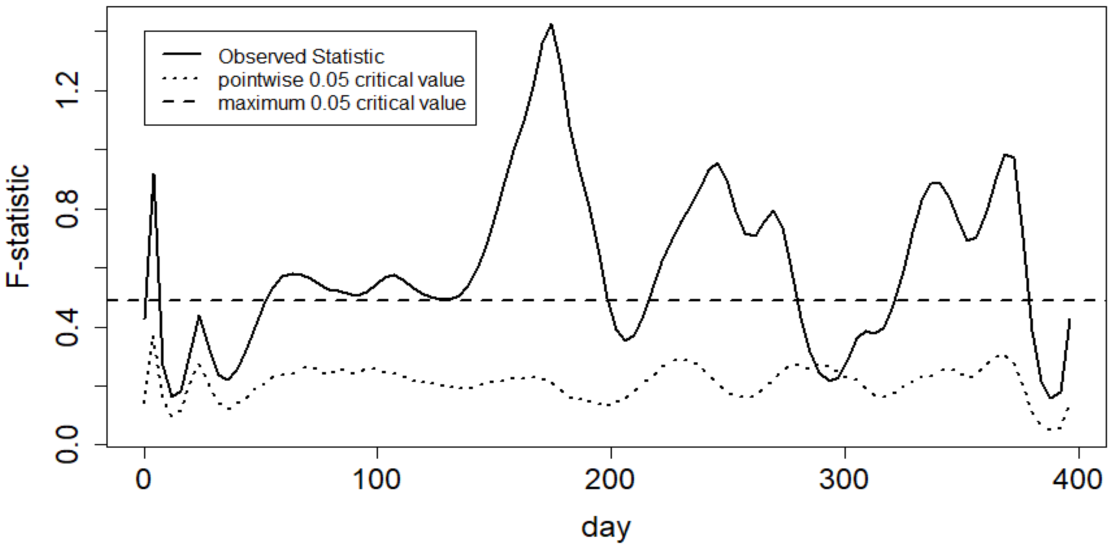

Figure 3.1.11-1, New COVID Cases for Functional 200 Permutation F-Test Statistic by Day

In this figure it shows the solid line as the observed statistic of the new COVID cases. If the observed statistic is below the dotted pointwise 0.05 critical value, then there is no regional effect or difference in that time period. The figure shows that the observed statistic is above the pointwise critical line which means that there is a difference in regions for the new COVID cases. Around day 300 is the only time that the observed statistics line falls below the pointwise value that there isn’t a difference in new COVID cases for the different regions. For most of this data the F statistic (Observed Statistic) is above the pointwise 0.05 critical value, this means that we can reject the null hypothesis and conclude that there is a difference in the regions for new COVID cases.

Figure 3.1.11-2, New COVID Cases for Functional 5,000 Permutations F-Test Statistic by Day with Bonferroni Correction, shows if there is a difference between the regions with the new COVID cases adjusting for the different sets of data.

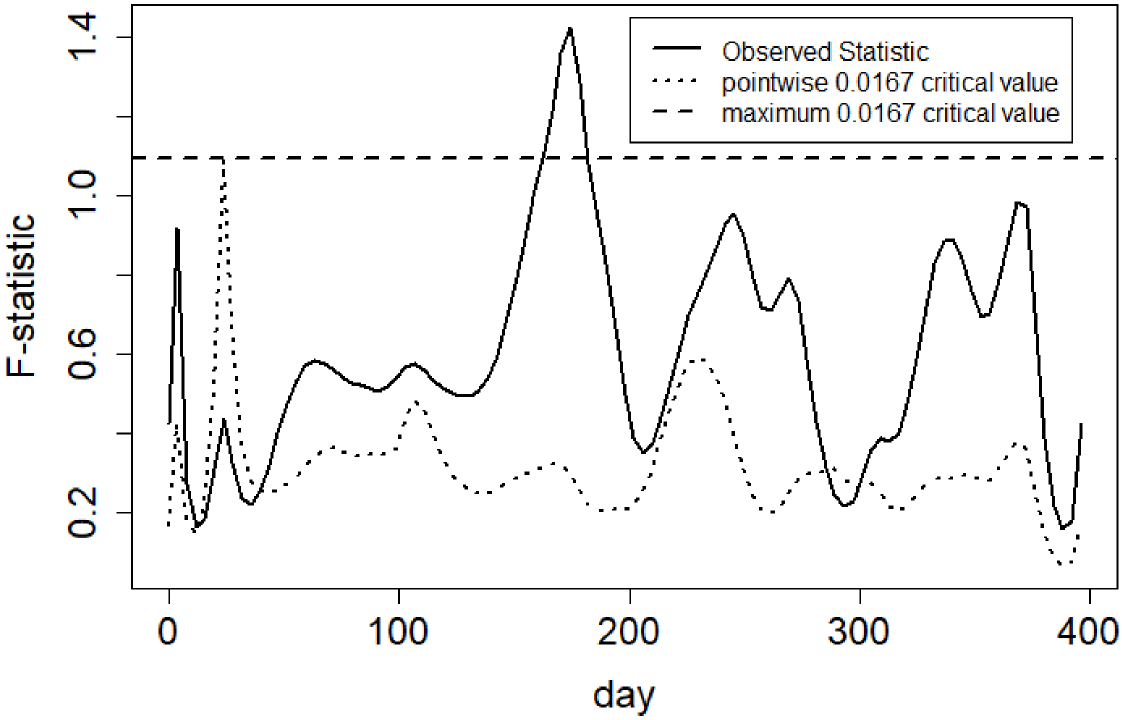

Figure 3.1.11-2, New COVID Cases for Functional 5,000 Permutations F-Test Statistic by Day with Bonferroni Correction

In this figure with the Bonferroni correction, it shows for all the data over time there is a difference in the regions for new COVID cases. In Figure 3.1.11-1, it showed around day 300 the data is below the pointwise 0.05 critical value. With the Bonferroni correction of the different data sets, the observed line is still below the 0.0167 critical line at that time and around days 30 to 60. Figure 3.1.11-2 shows the data with 5,000 permutations of the data. A permutation is putting the data in different arrangements of the data to get a more accurate output of the F-Test.

Figure 3.1.11-3 shows the data with 10,000 permutations to try to get the F-Test value to be more accurate. Comparing Figure 3.1.11-2 and Figure 3.1.11-3, the outputs for both look to be about the same. Since 5,000 and 10,000 permutations are the same this shows that using 5,000 permutations is accurate enough to use.

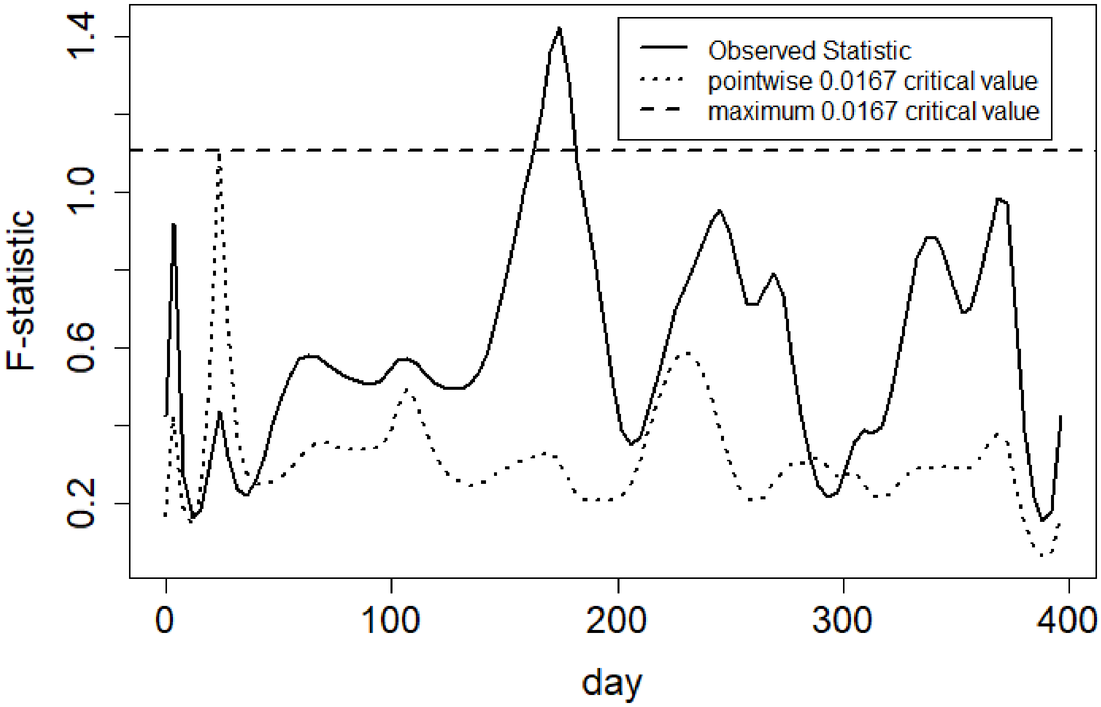

Figure 3.1.11-3, New COVID Cases for Functional 10,000 Permutations F-Test Statistic by Day with Bonferroni Correction

3.2 COVID Hospitalizations Analysis

Section 3.2 COVID Hospitalizations Analysis, looks at the COVID Hospitalized patients for each state and region by day. Future sections will also look at the contour and correlation coefficient plots to determine during which months the most COVID hospitalized patients are identified per region and state. This section, as well as future sections, will also look at the following:

- eigenvalues,
- cumulative percentage,
- functional principal component functions,
- beta graphs,
- regression coefficients,
- functional F-Test.

Analyzing the state and region data utilizing these different analysis methods will help compare the hospitalized patients by state and region then come to a valuable conclusion about the data.

For this data, the following are the null and alternative hypotheses::

- Null Hypothesis: H_0: There is no difference in the regions for COVID hospitalized patients
- Alternative Hypothesis: H_a: There is a difference in the regions for COVID hospitalized patients.

3.2.1 Hospitalizations by Region and State

The first data set that is analyzed is the COVID-19 Hospitalized Patients by each region and state. Figure 3.2.1-1, Hospitalized Patients per Region by Day displays for each region the number of hospitalized patients for that day by the population.

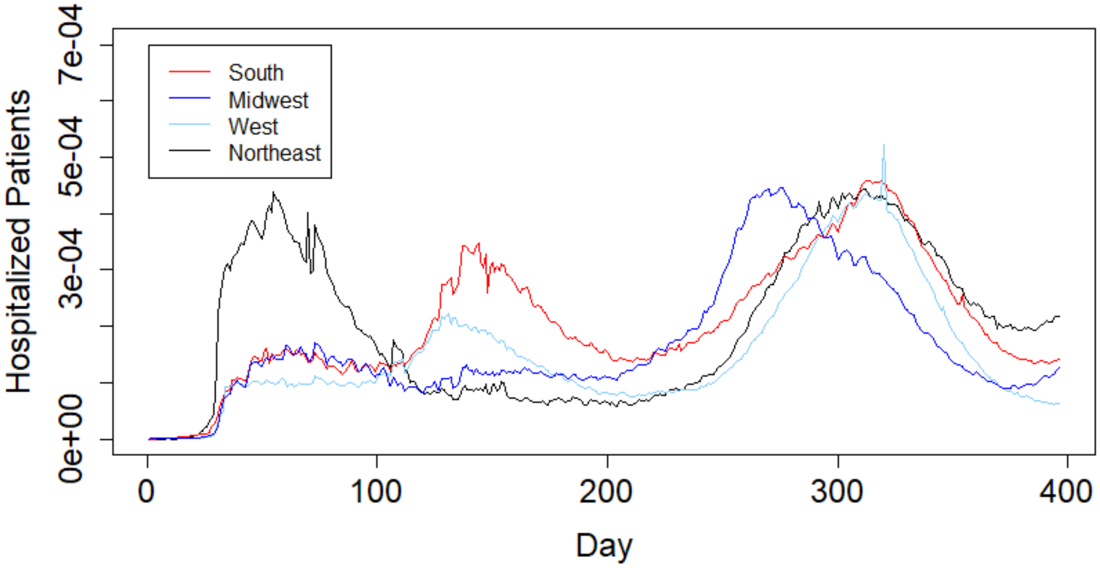

Figure 3.2.1-1, Hospitalized Patients per Region by Day

Figure 3.2.1-1 shows that the number of hospitalized patients starts low in all four regions, then in the Northeast region hospitalized patients spike up first. Looking at day 150 we see that the South and the West spike up and then go back down at day 200. Around day 250 all regions spike up and then go back down to be about the same constant number at day 400.

Figure 3.2.1-2, Hospitalized Patients per State by Day, displays the number of hospitalized COVID patients per day by each region by the color and then the different lines are the states.

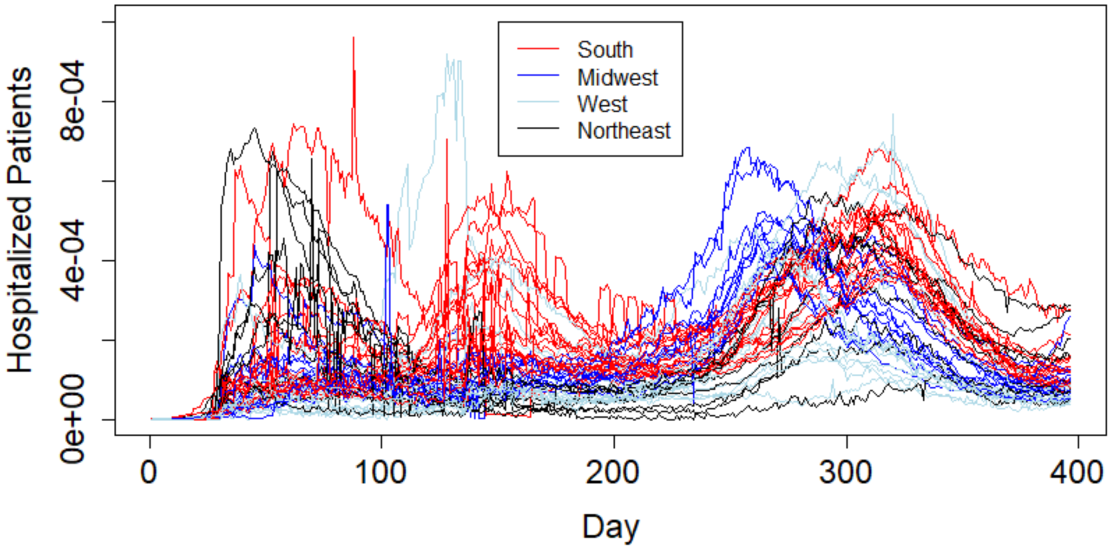

Figure 3.2.1-2, Hospitalized Patients per State by Day

When looking at Figure 3.2.1-2, some states between day 0 and 150 are a lot higher than the other states. After day 200 all of the lines on the graph seem to be following the same trend. There are a few states that stand out from the others in the beginning. In Figure 3.2.1-3 it will show which states are standing out from the other in the beginning of Figure 3.2.1-2.

Figure 3.2.1-3, Hospitalized Patients per Region by State by Day, displays the four regions and for each region, the states that are noticeable (have high hospitalized patients) in that region.

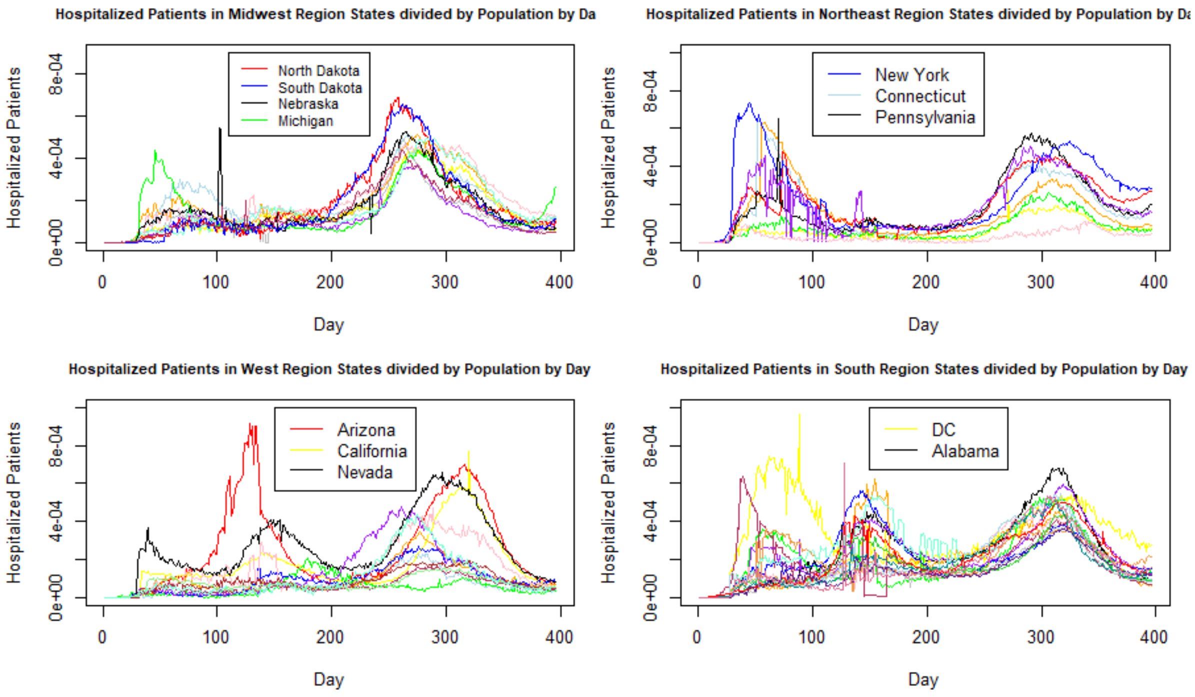

Figure 3.2.1-3, Hospitalized Patients per Region by State by Day

Looking at the Midwest Region States first, North Dakota, South Dakota, Nebraska, and Michigan have spikes and the highest number of COVID hospitalized patients. In the Northeast Region, New York, Connecticut, and Pennsylvania are the states that have some spikes or are the highest. In the West Region the states that stand out the most are Arizona, California, and Nevada. Arizona stands out the most among Western states, with a lot of new hospitalized patients around day 120. In the South region DC and Alabama stand out. There are some spikes in the data for some of the states. Most of the states in this region also follow the same pattern.

3.2.2 Smoothed Line of Hospitalizations

This section displays the smoothed lines of the hospitalized patient's data. In the beginning of section 3.1.2, it was analyzed that the best smoothing lambda to use is 1e4. For this section and data, lambda 1e4 will also be used to smooth the data.

Figure 3.2.2-1, Smoothed Line of Hospitalized Patients per Region by Day, illustrates why a lambda of 4 is selected for the data.

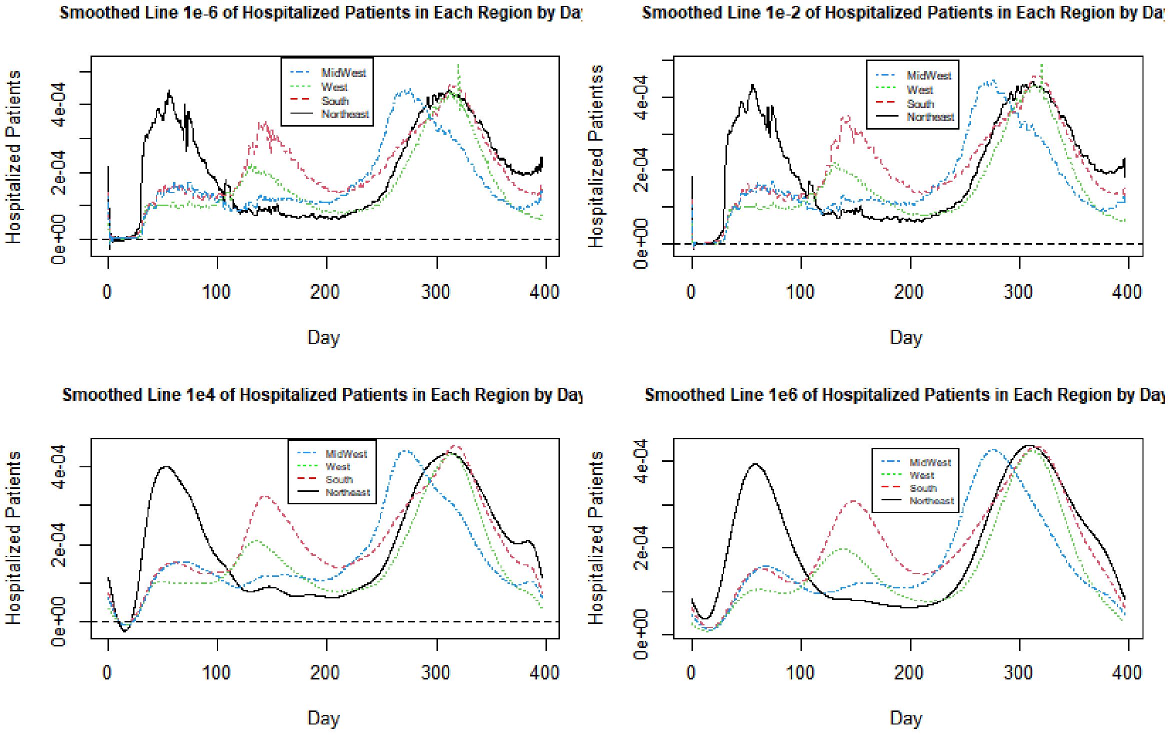

Figure 3.2.2-1, Smoothed Line of Hospitalized Patients per Region by Day

In Figure 3.2.2-1 there are four different smoothed line graphs to evaluate and determine which one is the best fit for this data. The first smoothed line graph is 10^{-6}. Here the graph shows lines still spike up and down with the data. The line needs to be more smoothed to be able to see the trend lines better. The second graph smooths the line out more to 10^{-2}, but still has some up and down spikes and could look better. The next graph is the smoothed line 10^{4}, this graph looks a lot better with the data and lines smoothed out more. The data is smoothed out one more time using a lambda of 10^{6}. For this lambda value, the graph is too smooth. The data needs to have some curves, to be able to see what is going on and to be able to analyze the data better. When looking at Figure 3.1.2-1 and Figure 3.2.2-1 the best lambda that will be utilized in this data analysis thesis is 1e4.

Figure 3.2.2-2, Smoothed Line of Hospitalized Patients per Region by Day, displays the smoothed lines of the hospitalized patients per Region.

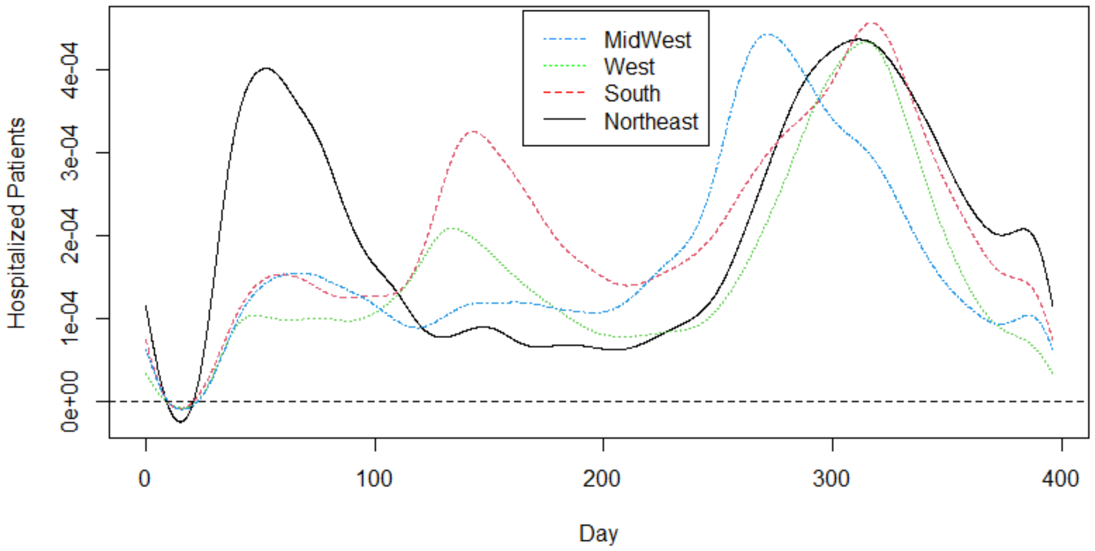

Figure 3.2.2-2, Smoothed Line of Hospitalized Patients per Region by Day

In Figure 3.2.2-2, with the smoothed line of the hospitalized patients per region by day, the Northeast area spikes up at day 50, then goes back down. Around day 150 the South region goes up higher than the other regions. Then around day 250 all regions spike up with all having about the same number of hospitalized patients.

Figure 3.2.2-3, Smoothed Line of Hospitalized Patients for States by Day, displays the number of hospitalized patients by states.

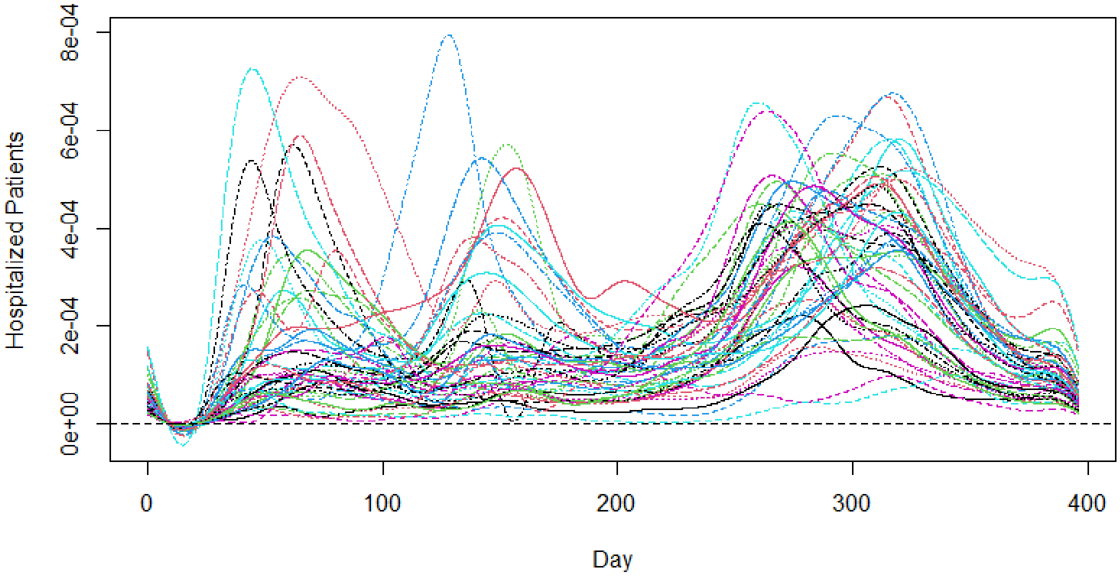

Figure 3.2.2-3, Smoothed Line of Hospitalized Patients for States by Day

In Figure 3.2.2-3, which is the smoothed line of the hospitalized patients for states by day, one state stands out more compared to the rest around day 120. The new hospitalized patients follow along the same paths and there aren’t many states that stand out from the others.

3.2.3 Smoothed Line Graph of Hospitalizations

This section looks at the hospitalized COVID patients over time and the average over time for all the data. Figure 3.2.3-1, Smoothed Line of Hospitalized Patients with the Mean Line per Region by Day, shows the number of hospitalized patients in that region for that day over time.

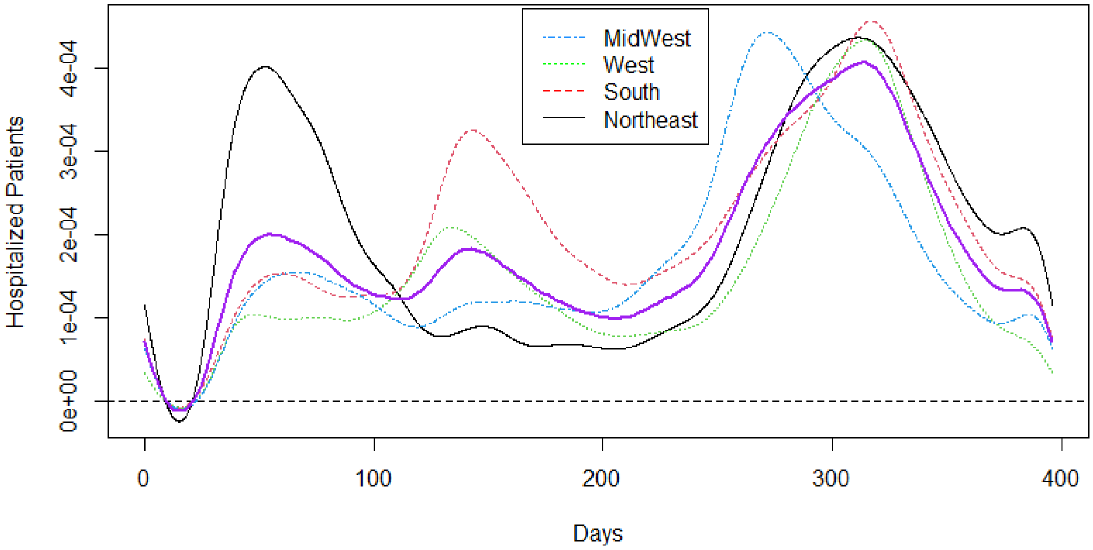

Figure 3.2.3-1, Smoothed Line of Hospitalized Patients with the Mean Line per Region by Day

The purple line shows the mean positive cases for all the regions over time for that day. The average line appears to be in the middle of the four regions, following the same trend of those Regions. Around day 50 the Northeast region pulls the average up to make it higher than the other regions. Around day 150 the South is up a lot higher than the other regions. At day 300, all regions are very high, as well as the average line.

Figure 3.2.3-2, Smoothed Line of Hospitalized Patients with the Mean Line for States by Day, shows the smoothed lines of the COVID hospitalized patients. On this graph, the mean line of the data is plotted in a solid bold purple line as well as the hospitalized patients by states lines. The mean line of the data helps determine which states stand out from the average.

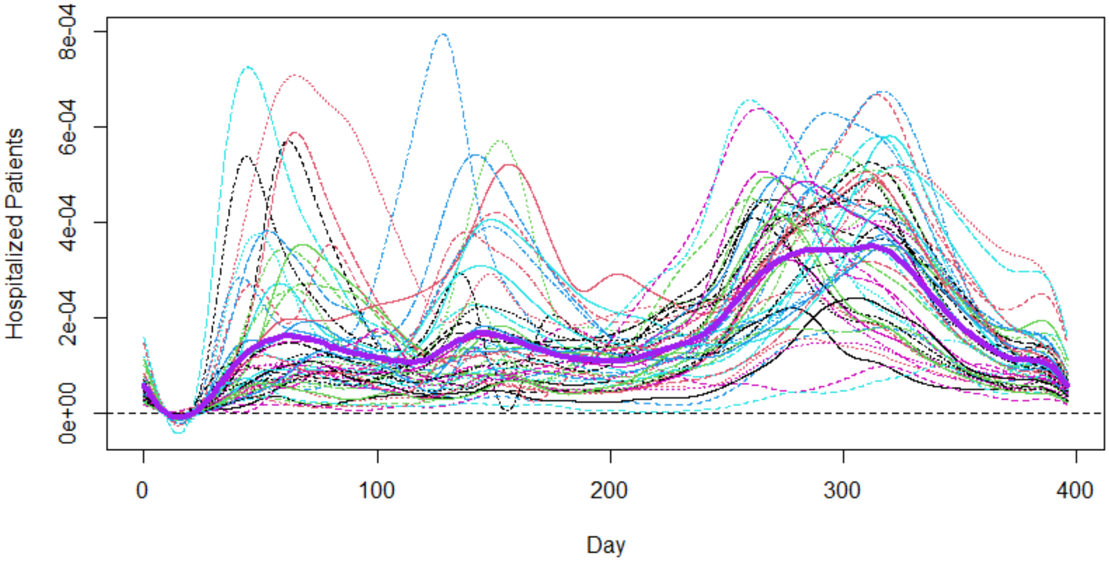

Figure 3.2.3-2, Smoothed Line of Hospitalized Patients with the Mean Line for States by Day

Figure 3.2.3-2 shows that around days 50, 110, and 250 to 320 there are spikes in the data. For those days the mean line goes up but not as high as some of the data does. There are some states that pull the mean line up, compared to the other states that have a lower hospitalized patient count.

The next figure displays the smoothed lines of the data, minus the mean line. By plotting the results, Figure 3.2.3-3 displays which regions are above or below the average of the data.

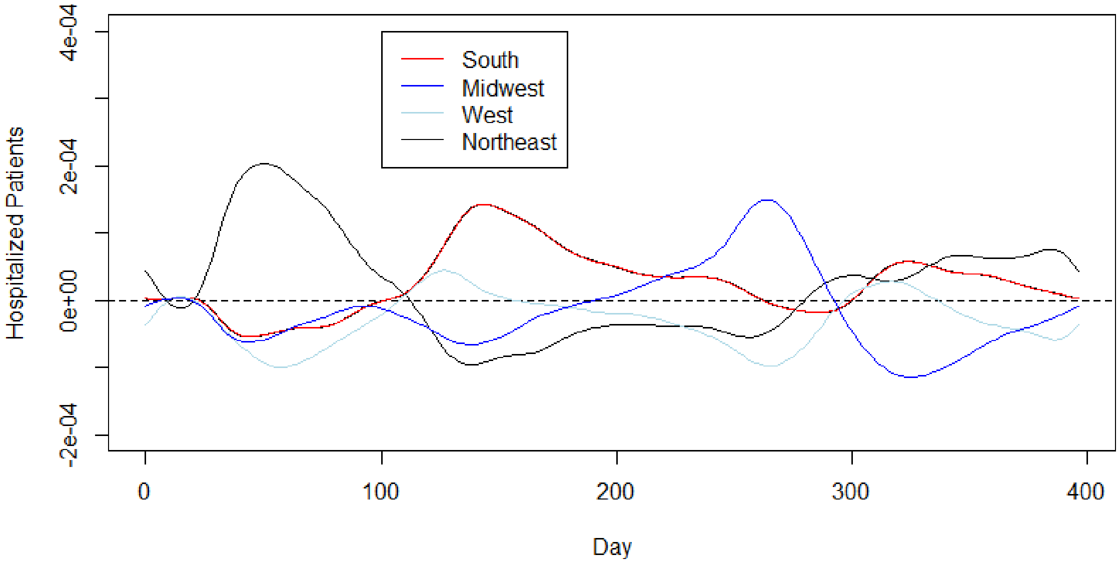

Figure 3.2.3-3, Smoothed Line of Hospitalized Patients minus the Mean per Region by Day

Figure 3.2.3-3 shows the smothered lines of the hospitalized patients for each region minus the mean line. This figure shows that the Northeast area is again the highest in the beginning like is shown in Figure 3.2.3-1. The South region is higher at first, then drops down and comes back up at the end. The West stays constant going up and down over time but nothing to concerning. The Midwest stands out the most around day 250, where the hospitalizations spike up higher than the other regions.

The last Figure 3.2.3-4, Smoothed Line of Hospitalized Patients minus the Mean for States by Day, shows the number of cases in each region by the states minus the averages.

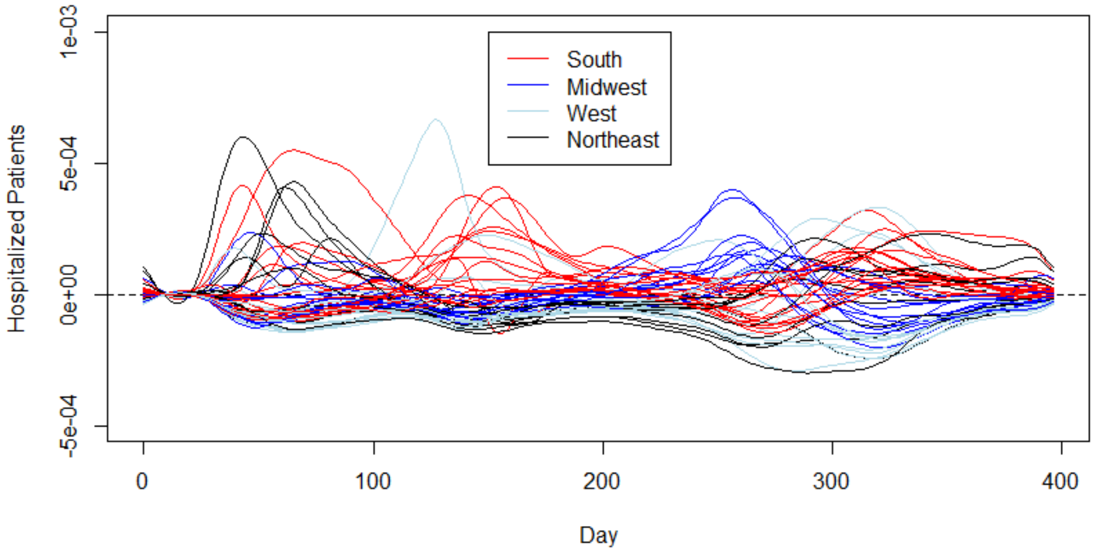

Figure 3.2.3-4, Smoothed Line of Hospitalized Patients minus the Mean for States by Day

Looking at the last Figure 3.2.3-4, in the beginning there are a few Northeast and South states that are above the average that stand out. Around day 110 it switches, and a state in the West Region stands out above the average as well as the other states. Then around day 150 the South states are the highest and really stand out above the rest. Around day 250, the Midwest states stand out from the other states. From there as time goes on the states flatten out to be about the same around the 0 and the average. Over the time period, each of the four different regions stand out at least one time, from the other regions.

3.2.4 Variance-Covariance of Hospitalizations

In Figure 3.2.4-1, Contour Plot of Hospitalized Patients per Region by Day, shows the contour plot of the variance-covariance of the COVID hospitalized patient cases per region by day.

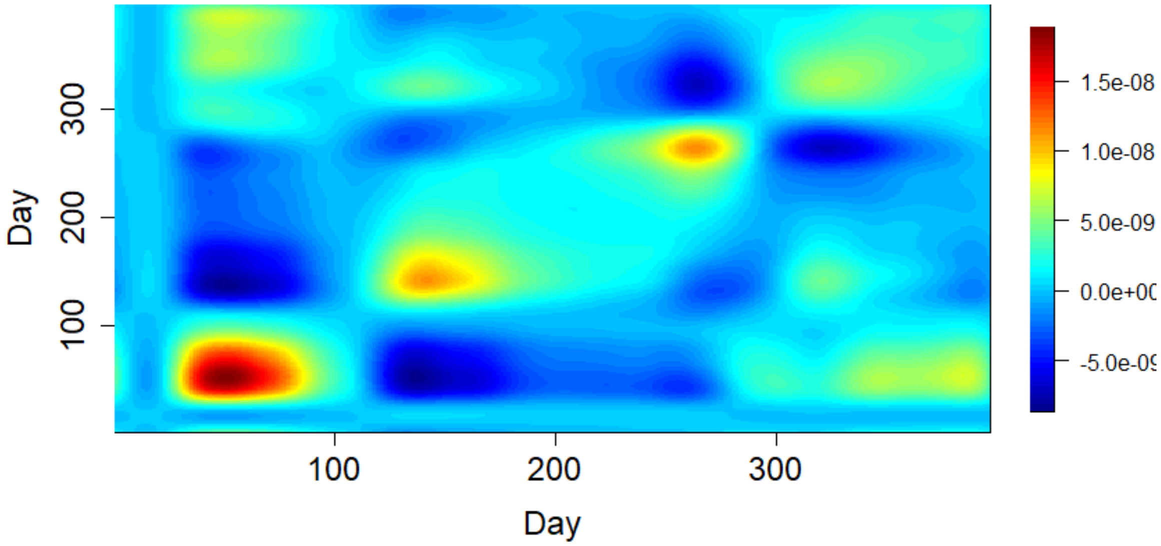

Figure 3.2.4-1, Contour Plot of Hospitalized Patients per Region by Day

This plot shows the covariance surface of the hospitalized patients per region by day. The diagonal of the plot is also the variance of the data. On the diagonal of the plot, for each day the variance counts are a mirror image of each other. The dark red area on the plot is the highest hospitalized patients which is 1.5e-08, and is along the diagonal which is part of the variance.

Figure 3.2.4-2, Contour Plot of Hospitalized Patients for States by Day, displays a contour plot of the variance-covariance of the COVID hospitalized patients for states by day.

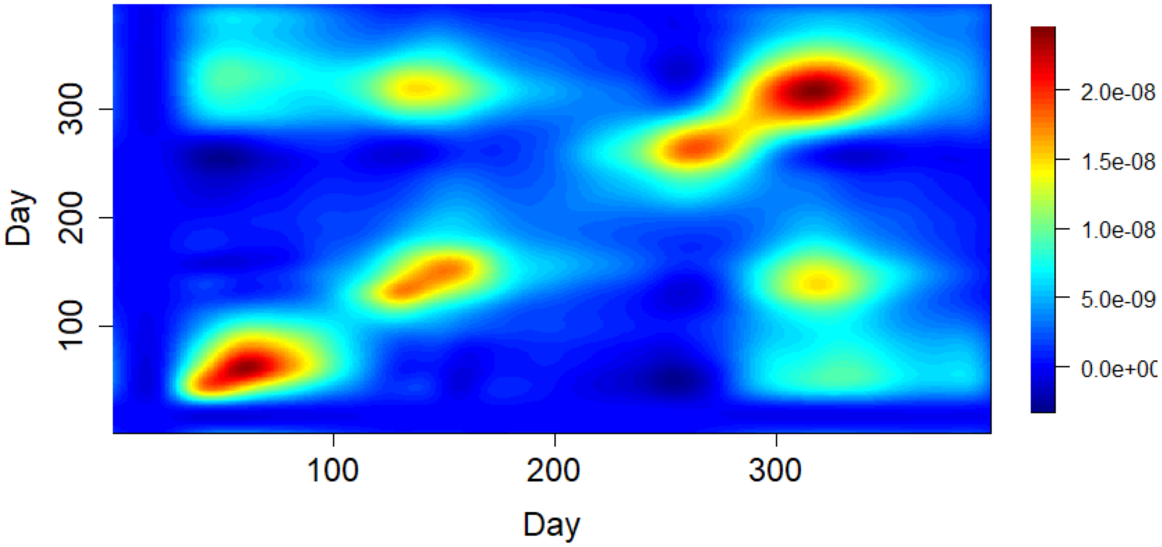

Figure 3.2.4-2, Contour Plot of Hospitalized Patients for States by Day

This plot shows the covariance surface of the Hospitalized Patients for states by each day. The diagonal of the plot is the variance of the data. Across the diagonal of the plot each side mimics the other. The next section will look at the correlation coefficient plots which are easier to read and describe the data.

3.2.5 Correlation Coefficient Plot Hospitalizations

This section investigates the correlation coefficient plots of the new hospitalized COVID patients over time. The X and the Y axis are both days to compare the time of year and the number of new hospitalized COVID patients there have been. Figure 3.2.5-1, Correlation Coefficient Plot of Hospitalized Patients per Region by Day, displays the correlation coefficient plot of the new hospitalized COVID patients per region by day.

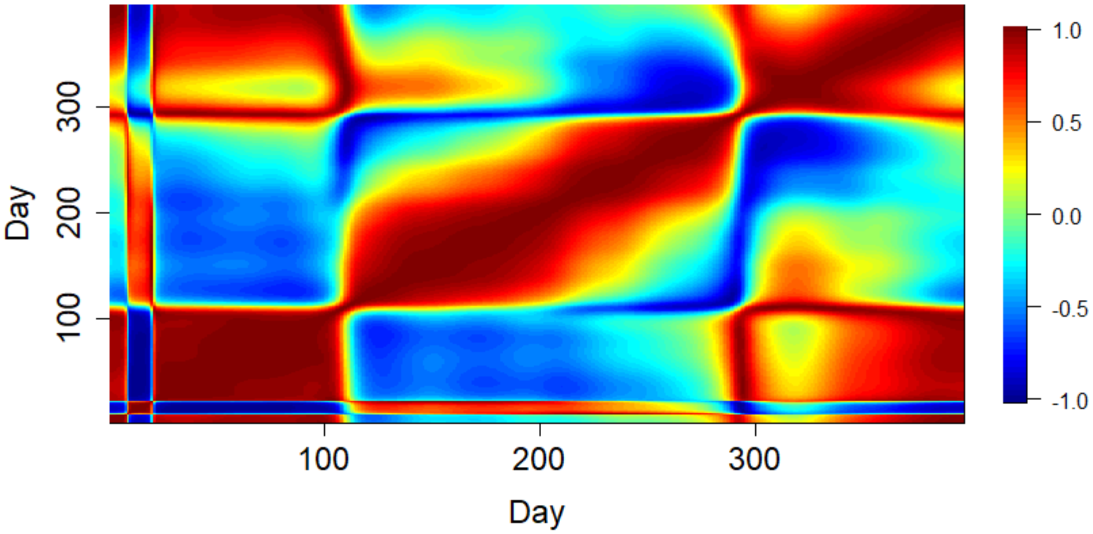

Figure 3.2.5-1, Correlation Coefficient Plot of Hospitalized Patients per Region by Day

Figure 3.2.5-1 is easier to read and conveys more than the variance-covariance plot that is in Figure 3.2.4-1, Contour Plot of Hospitalized Patients per Region by Day, above. Figure 3.2.5-1, along the diagonal, is equal to the value 1 which is the maximum correlation. The highest value is along the diagonal when it is 1, the lowest correlation is the darker blue (-1.0) which is the correlation between

June to July (x-axis) and April to June (y-axis) in the region areas. There is also a small spot of dark blue which is the correlation between March (x-axis) to March (y-axis).

Figure 3.2.5-2, Correlation Coefficient Plot of Hospitalized Patients for States by Day, is the correlation coefficient plot of the new hospitalized COVID patients for states by day. Along the diagonal is again, the highest correlation value of 1. The smallest correlation is in the dark blue again of -0.5, which is the correlation from March to March in the states areas.

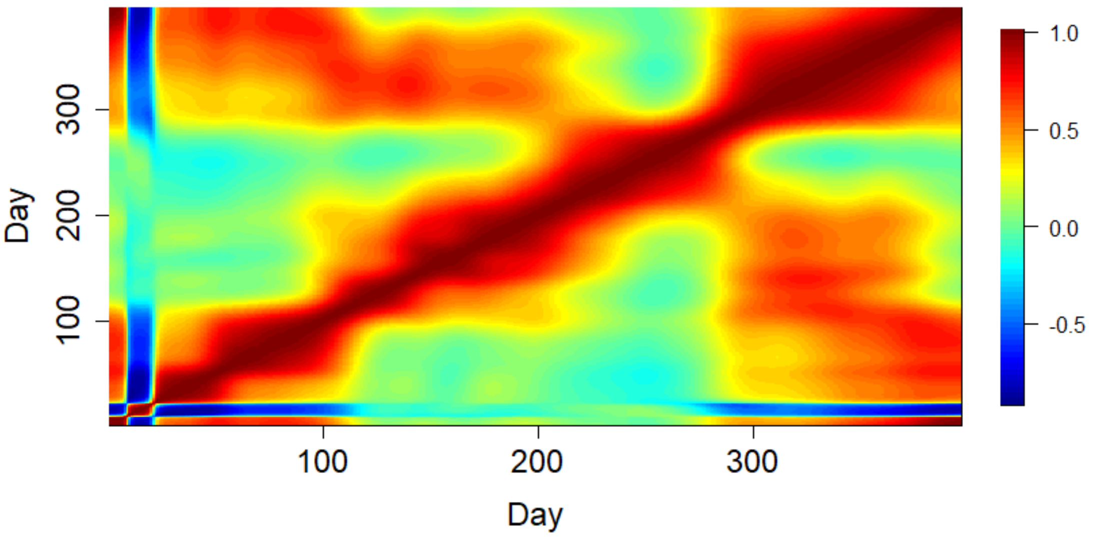

Figure 3.2.5-2, Correlation Coefficient Plot of Hospitalized Patients for States by Day

Figure 3.2.5-2 has brighter colors (green, yellow, orange, and red) than the darker blue colors. This shows that there is more of a correlation in the hospitalized patients over time than what there is in the new COVID cases over time (Figure 3.1.5-2).

3.2.6 Eigenvalue Plot Hospitalizations

This section examines the Functional Principal Component Analysis (FPCA) and the eigenvalues of the new hospitalized COVID patients. All the data sets use four Functional Principal Components (FPC) which was discussed in section 3.1.8. Figure 3.2.6-1, Eigenvalue Plot of Hospitalized Patients per Region by Day, displays each eigenvalue for each FPCA.

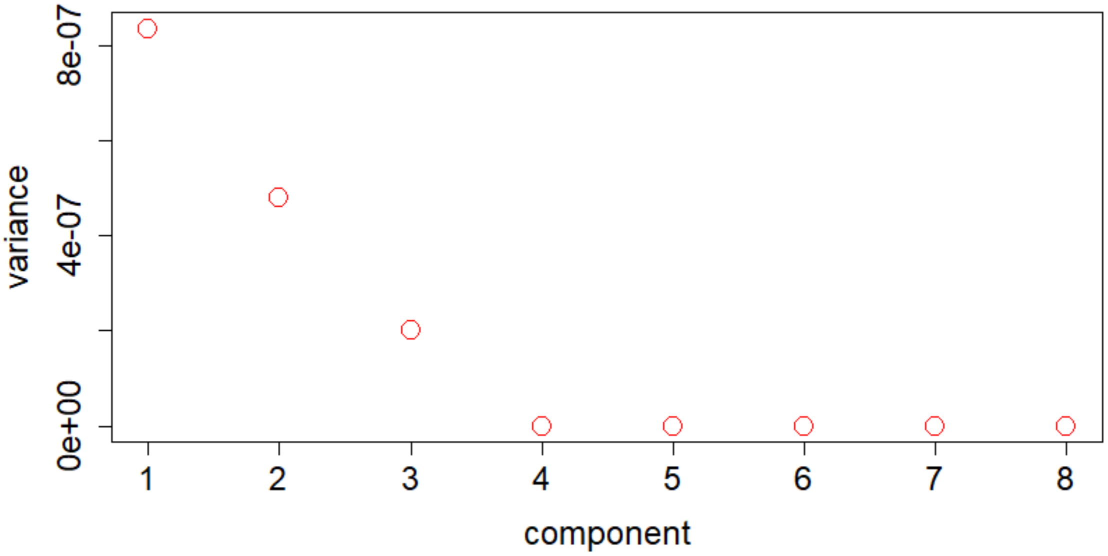

Figure 3.2.6-1, Eigenvalue Plot of Hospitalized Patients per Region by Day

Each eigenvalue is equal to the variance of the FPC score. As the eigenvalue component increases the variance value decreases and gets closer and closer to 0. The first eigenvalue is large and then the next one drops in value very fast.

Figure 3.2.6-2, Eigenvalue Plot of Hospitalized Patients for States by Day, examines the same values but for the hospitalized patients for states by day.

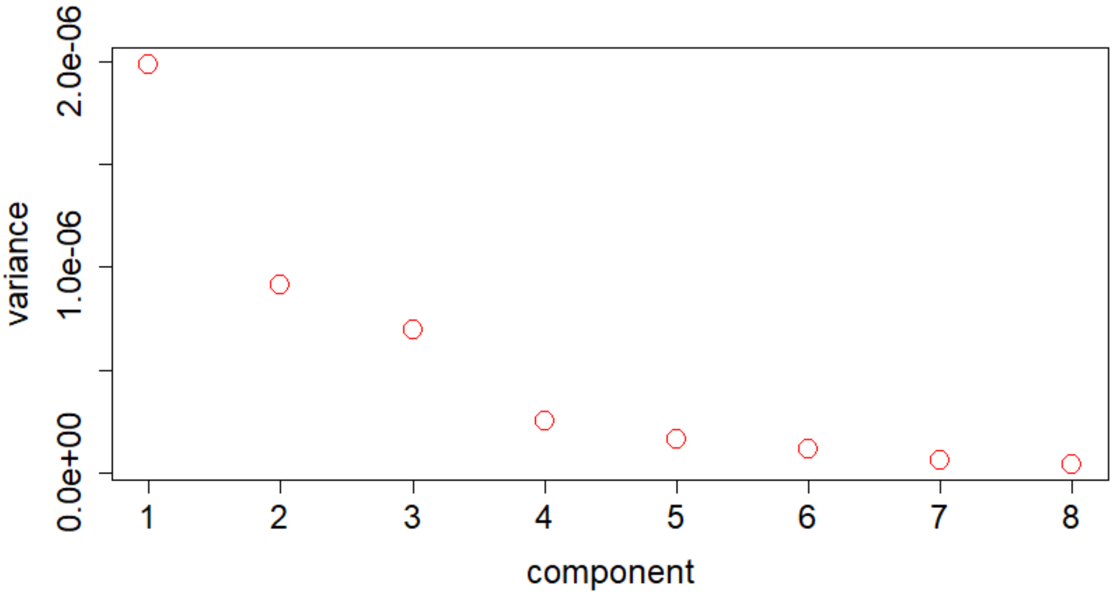

Figure 3.2.6-2, Eigenvalue Plot of Hospitalized Patients for States by Day

The values and eigenvalues for the new hospitalized COVID patients for states by day are bigger than the ones for new hospitalized COVID patients per region. The eigenvalues for Figure 3.2.6-2 drop more slowly than in Figure 3.2.6-1. The eigenvalues don’t seem to get to around 0 until the 5th eigenvalue.

3.2.7 Cumulative Percentage Plot Hospitalizations

This section looks at the cumulative percentage explained for the total variation plots of the hospitalized COVID patients. Theses plots are the sum of the ten eigenvalues divided by the sum of all the eigenvalues. This shows what number of FPC's (x-axis) are needed to show what percentage of the data (y-axis).

Figure 3.2.7-1, Cumulative Percentage Plot of Hospitalized Patients per Region by Day, displays the cumulative percentage of the hospitalized COVID patients per region by day.

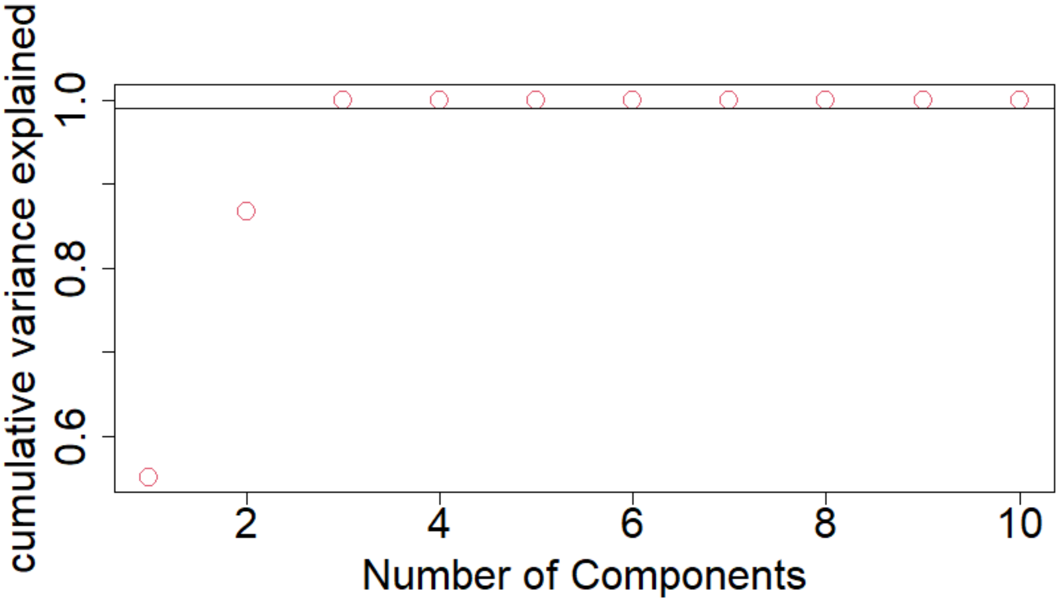

Figure 3.2.7-1, Cumulative Percentage Plot of Hospitalized Patients per Region by Day

Figure 3.2.7-1 shows that the first 3 FPCS explain about 99% of the total variation of the data. Once the graph reaches the 3rd and above FPCS the graph shows that the value is above the 99% line and remains about the same value. With this data the first 3 FPCs are needed to explain most of the data.

Figure 3.2.7-2, Cumulative Percentage Plot of Hospitalized Patients for States by Day, displays the cumulative percentage of the new hospitalized COVID patients for states by day.

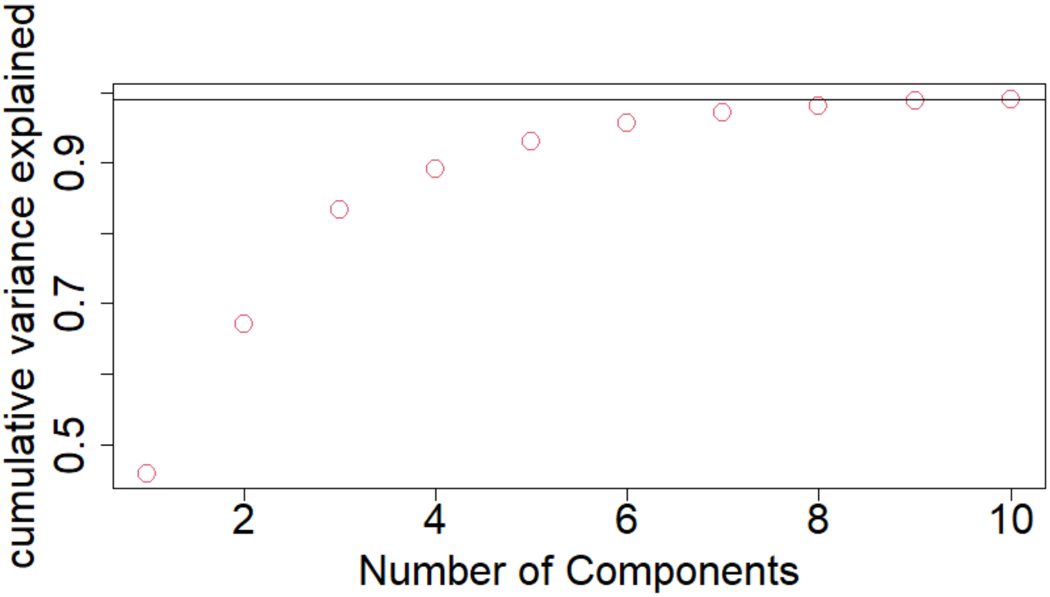

Figure 3.2.7-2, Cumulative Percentage Plot of Hospitalized Patients for States by Day

Figure 3.2.7-2 displays that more FPCs are needed to explain the data than in Figure 3.2.7-1. For the new hospitalized COVID patients for states by day, the number of components does not go above the 1 line until about 10 components. This data would need about 6 or 7 FPCs to explain about 99% of the data.

3.2.8 Functional Principal Component Functions Graph of Hospitalizations

Section 3.2.7 presented cumulative percentages and how many FPCs are needed to explain the data. Looking at Figures 3.2.7-1 shows that 3 are needed and 3.2.7-2 shows that 6 or 7 are needed. As discussed in section 3.1.8, 4 FPCs are used to stay consistent and to look further into the data.

Figure 3.2.8-1, Functional Principal Component Functions of Hospitalized Patients per Region by Day, looks deeper into explaining the data.

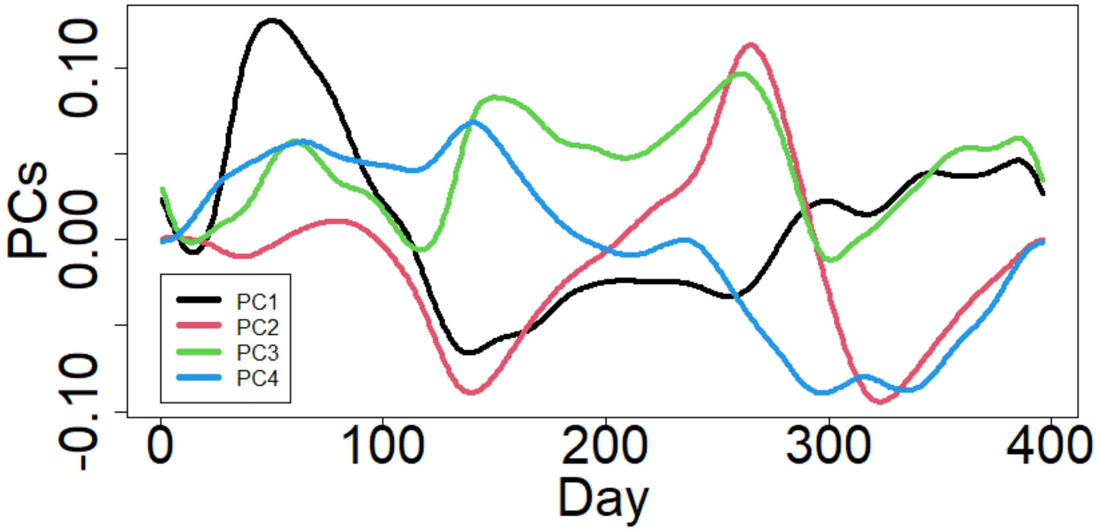

Figure 3.2.8-1, Functional Principal Component Functions of Hospitalized Patients per Region by Day

In Figure 3.2.8-1, PC1 matches up with the Northeast region, both spike up the highest around day 50. PC2, the red line, matches up the best with the Midwest region, staying down low most of the time until spiking up around day 300. PC3, the green line, matches up best with the south region, continuously going up and down and never really staying constant. The last PC4 then fits best with the west region, spiking up around day 150 and then staying lower than the other regions.

Figure 3.2.8-2, Functional Principal Component Functions of Hospitalized Patients for States by Day, displays the four Principal Components that best represents/explains the data.

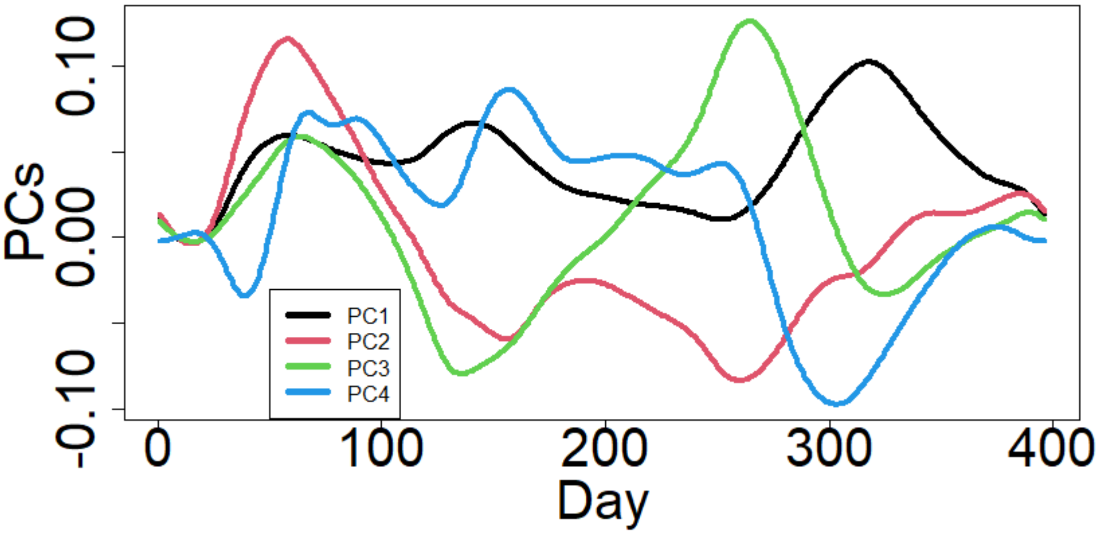

Figure 3.2.8-2, Functional Principal Component Functions of Hospitalized Patients for States by Day

Figure 3.2.8-2 is similar to Figure 3.2.8-1 (which represents Region Data), where Figure 3.2.8-2 represents data for each state and not just the regions. Referencing Figure 3.2.1-3, Hospitalized Patients per Region by State by Day, some of the states that stand out match up with some of the Principal Components (PC) in Figure 3.2.8-2. Alabama and Nevada match up best with PC1, Michigan, New York, and DC with PC2, Pennsylvania, North Dakota, and South Dakota with PC3, and

Arizona and Nebraska with PC4. All of the states will match up with one of the PC’s but those examples just show some of the states that stood out from the others.

3.2.9 FPC1 vs FPC2 Graphs of Hospitalizations

This section investigates the FPC1 on the x-axis and the FPC2 on the y-axis. This shows what states and regions are grouped together and are similar and which ones stand out from the others. Figure 3.2.9-1, FPC1 vs FPC2 Scores of Hospitalized Patients per Region by Day, displays FPC1 (x-axis) against FPC2 (y-axis).

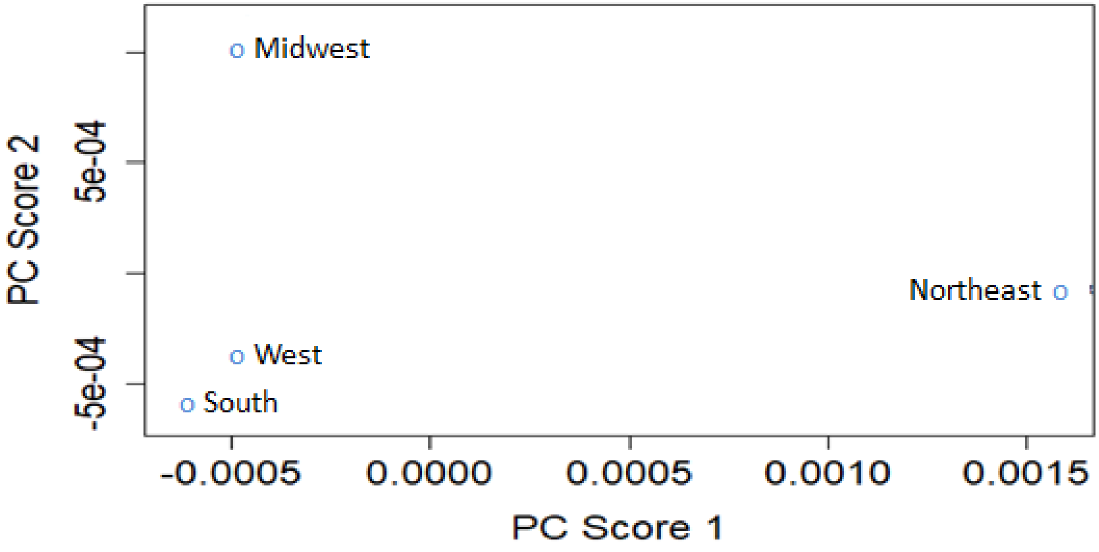

Figure 3.2.9-1, FPC1 vs FPC2 Scores of Hospitalized Patients per Region by Day

Looking at Figure 3.2.9-1 one can see that the South and the West regions are grouped together. With those regions grouped together, it shows that those two regions are similar with new hospitalized COVID patients. The Northeast and the Midwest are not grouped with any other region, showing that these regions aren't like the other regions in the new hospitalized COVID patients. Figure 3.2.9-2 looks at the same type of graph but with the different states. This shows what states are similar in the new hospitalized COVID patients. It also shows which states are outliers and aren't similar to the other states.

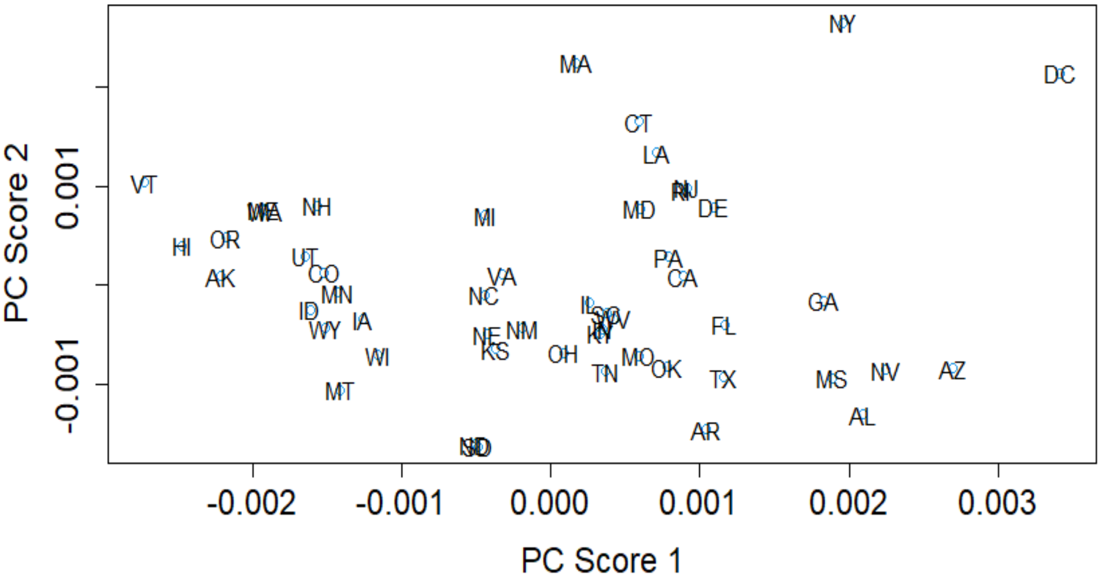

Figure 3.2.9-2, FPC1 vs FPC2 Scores of Hospitalized Patients for States by Day

Figure 3.2.9-2, most of the states are grouped together but there are a few that stand out from the others. DC and New York are off to the top right from the other states, this shows that those states didn't follow along with the other states in the new hospitalized COVID patients. Massachusetts is one state that is a little higher than the other states and stands out a little more.

3.2.10 Regression Coefficients

Figure 3.2.10-1, Hospitalized Patients Mean, Difference and Predicted by Region by Day, displays the regression coefficients and the differences between the mean and each region hospitalized COVID patients.

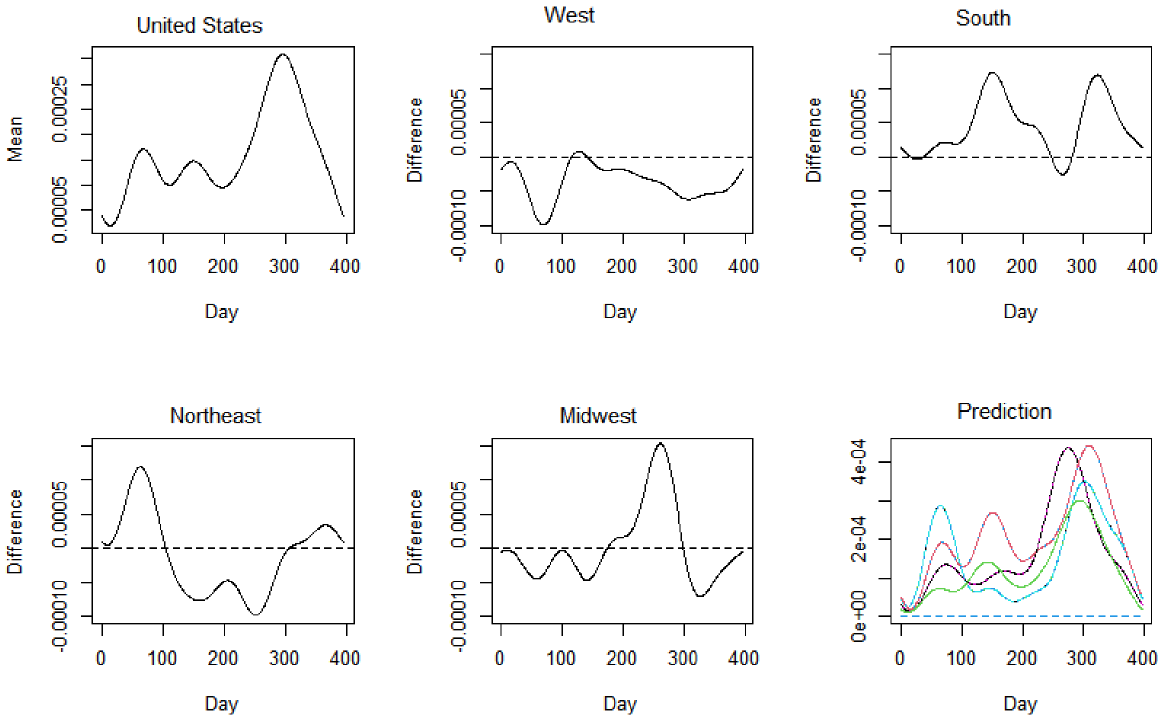

Figure 3.2.10-1, Hospitalized Patients Mean, Difference and Predicted by Region by Day

The United States graph in Figure 3.2.10-1 shows the mean of all the hospitalized COVID patients over the time frame. The next four graphs show the difference in that area of the new hospitalized COVID patients and the mean from the United States. With the four regions, when they are in the positive values (above the 0 dotted line) it shows when they have the highest impacts. When the values are in the positive it shows that they had more hospitalized COVID patients in that region and what the mean is for that time. The graph that stands out the most is the south region. It is in the positive values the most which means that area had the most hospitalized COVID patients. The last graph shows the predictions or the positive hospitalized COVID patients for that region.

3.2.11 Functional F-Test

This last section looks at the Functional F-Test for the new hospitalized COVID patients over time. As was noted earlier, for this data, consider the hypotheses:

- Null Hypothesis: H_0: There is no difference in the regions for hospitalized patients
- Alternative Hypothesis: H_a: There is a difference in the regions for hospitalized patients

Figure 3.2.11-1, Hospitalized Patients for Functional F-Test Statistic by Day, shows if there is a difference between the regions with the new hospitalized COVID patients.

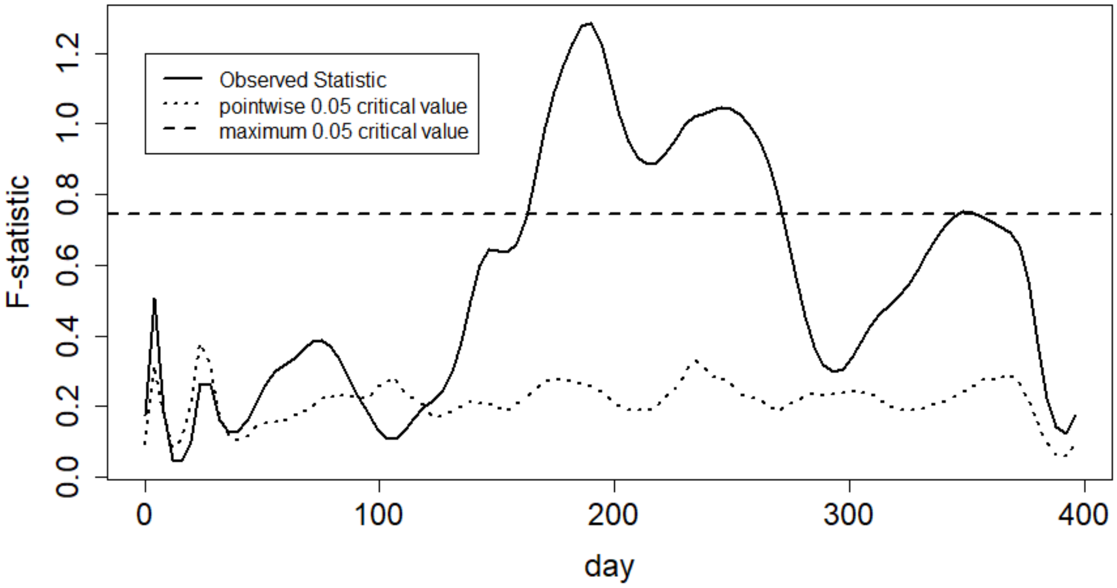

Figure 3.2.11-1, Hospitalized Patients for Functional Permutation F-Test Statistic by Day

In this figure it shows the solid line as the observed statistic of the new hospitalized COVID patients. If the observed statistic is below the dotted pointwise 0.05 critical value, then there is no regional effect or difference in that time period. The figure shows that the observed statistic is above the pointwise critical line which means that there is a difference in regions for the new hospitalized COVID patients. Around day 30 to 50 and 90 to 120 are the only times that the observed statistics line falls below the pointwise value that there isn’t a difference in new hospitalized COVID patients for the different regions. For most of this data the F statistic (Observed Statistic) is above the pointwise 0.05 critical value. This means that we can reject the null hypothesis and conclude that there is a difference in the regions for new hospitalized COVID patients.

Figure 3.2.11-2, Hospitalized Patients for Functional 5,000 F-Test Statistic by Day with Bonferroni Correction, shows if there is a difference between the regions of hospitalized patients adjusting for the different sets of data.

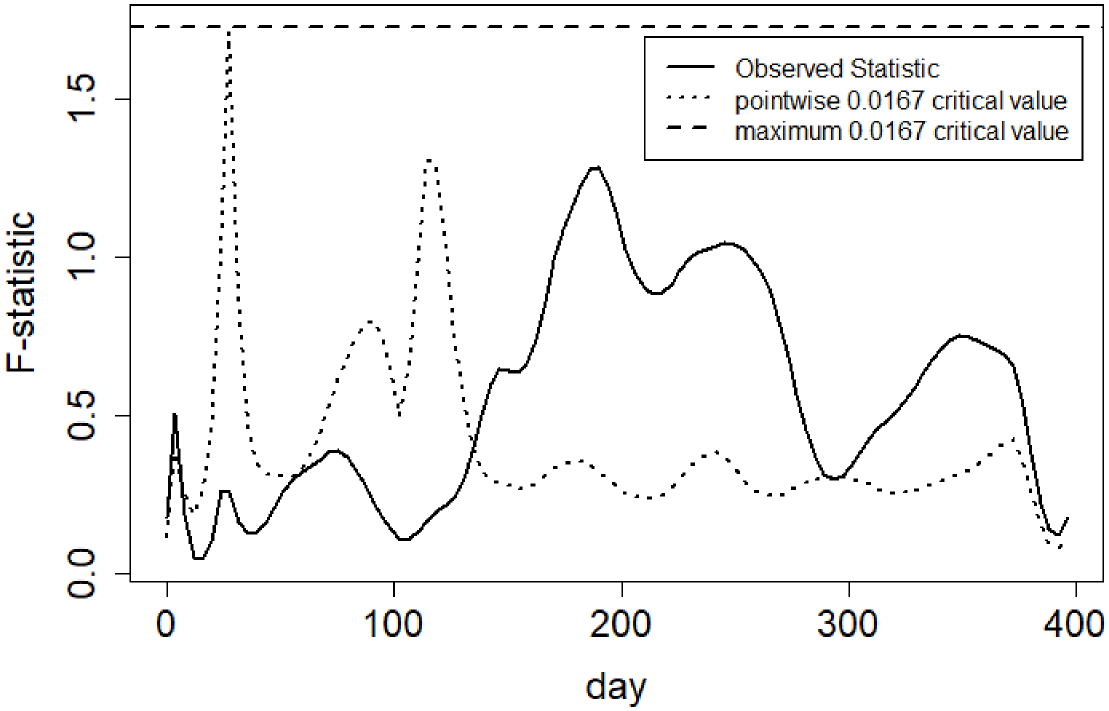

Figure 3.2.11-2, Hospitalized Patients for Functional 5,000 Permutation F-Test Statistic by Day with Bonferroni Correction

In this figure with the Bonferroni correction, it shows for all the data over time there is a difference in the regions for hospitalized COVID patients. In Figure 3.2.11-1, it shows around day 30 to 50 and 90 to 120 the data is below the pointwise 0.05 critical value. With the Bonferroni correction in the F-test the data is a lot more sensitive now. With the data being more sensitive the observed line falls below the pointwise line more often than it did without the Bonferroni adjustment. COVID patients. Figure 3.2.11-2 shows the data with 5,000 permutations of the data. A permutation is putting the data in different arrangements of the data to get a more accurate output of the F-Test.

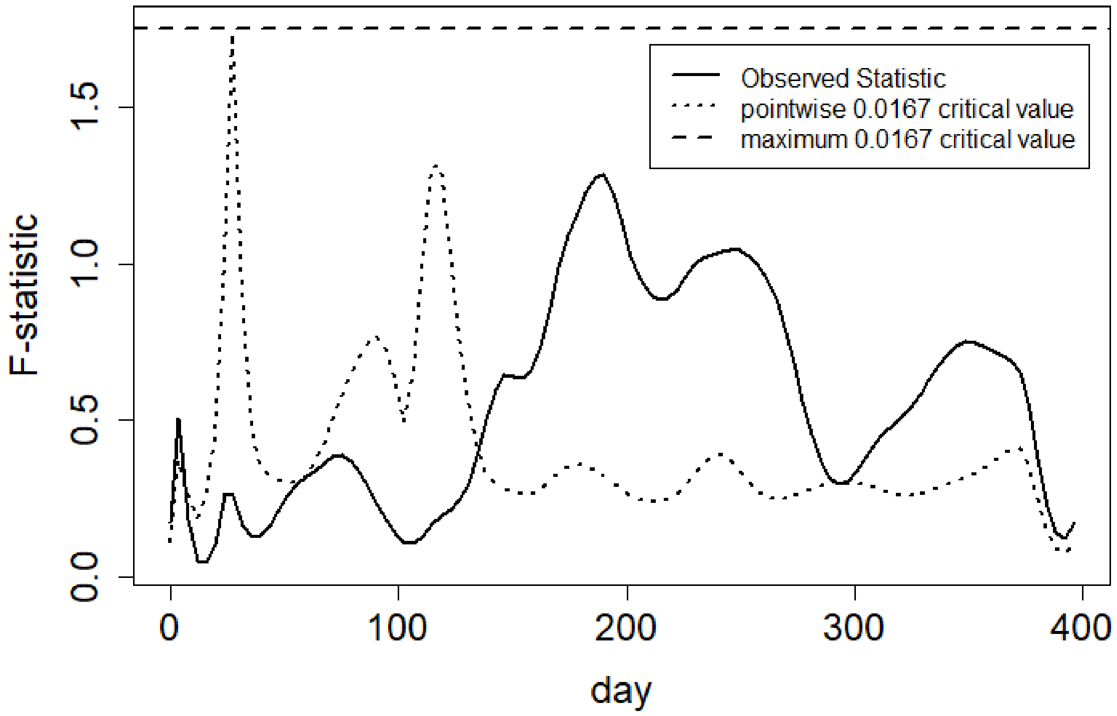

Figure 3.2.11-3, Hospitalized Patients for Functional 10,000 Permutation F-Test Statistic by Day with Bonferroni Correction

Figure 3.2.11-3 shows the data with 10,000 permutations to try to get the F-Test value to be more accurate. Comparing Figure 3.2.11-2 and Figure 3.2.11-3, the outputs for both look to be about the same. Since 5,000 and 10,000 permutations are the same this shows that using 5,000 permutations is accurate enough to use.

3.3 New COVID Deaths Analysis

Section 3.3, New COVID Deaths Analysis, looks at the new COVID deaths for each state and region by day. Later sections will also look at the contour and correlation coefficient plots to determine during which months the most new COVID deaths are identified per region and state. This section, as well as later sections, will also look at the following:

- eigenvalues,
- cumulative percentage,
- functional principal component functions,
- beta graphs,
- regression coefficients,
- functional F-Test.

Analyzing the state and region data utilizing these different analysis methods will help compare the new deaths by state and region and then come to a valuable conclusion about the data.

For this data, the following are the null and alternative hypotheses:

- Null Hypothesis: H_0: There is no difference in the regions for new COVID deaths
- Alternative Hypothesis: H_a: There is a difference in the regions for new COVID deaths

3.3.1 New Deaths by Region and State

In Figure 3.3.1-1, New COVID Deaths per Region by Day, displays for each region the number of new COVID deaths for that day by the population.

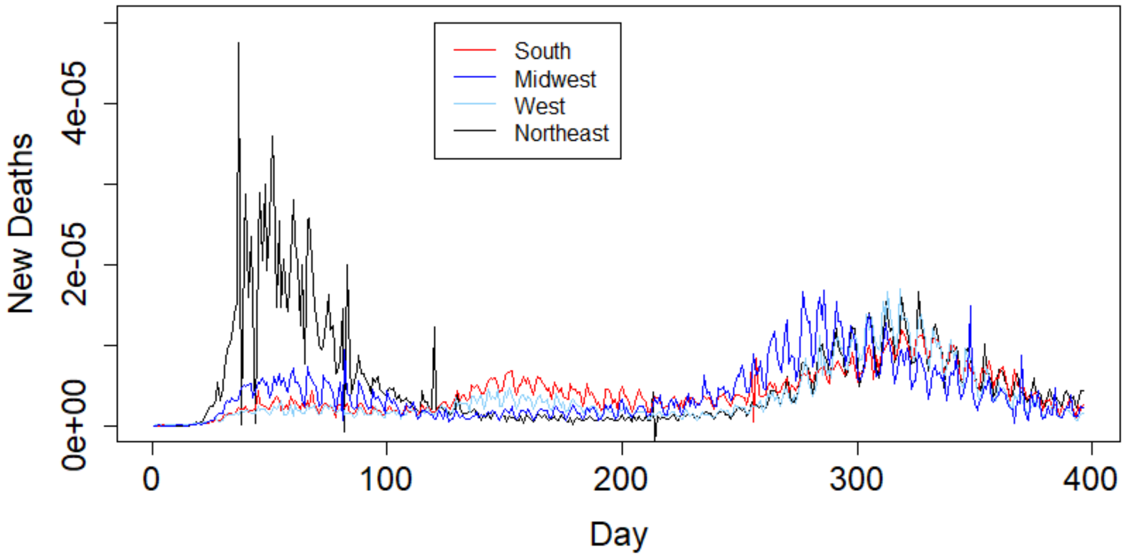

Figure 3.3.1-1, New COVID Deaths per Region by Day

Figure 3.3.1-1 shows that the number of new deaths starts low in all four regions, then in the Northeast region new deaths spike up first. Looking at day 120, it shows that the Northeast spikes up a little and then goes back down. Around day 250 all regions start to go up slowly to the max around day 300. All regions then start to go back down to be about the same constant number at day 400. The Northeast region has the highest new COVID deaths around day 50 comparing all the regions.

Figure 3.3.1-2, New COVID Deaths for States by Day, displays the number of new deaths per day by each region by the color with different lines for individual states.

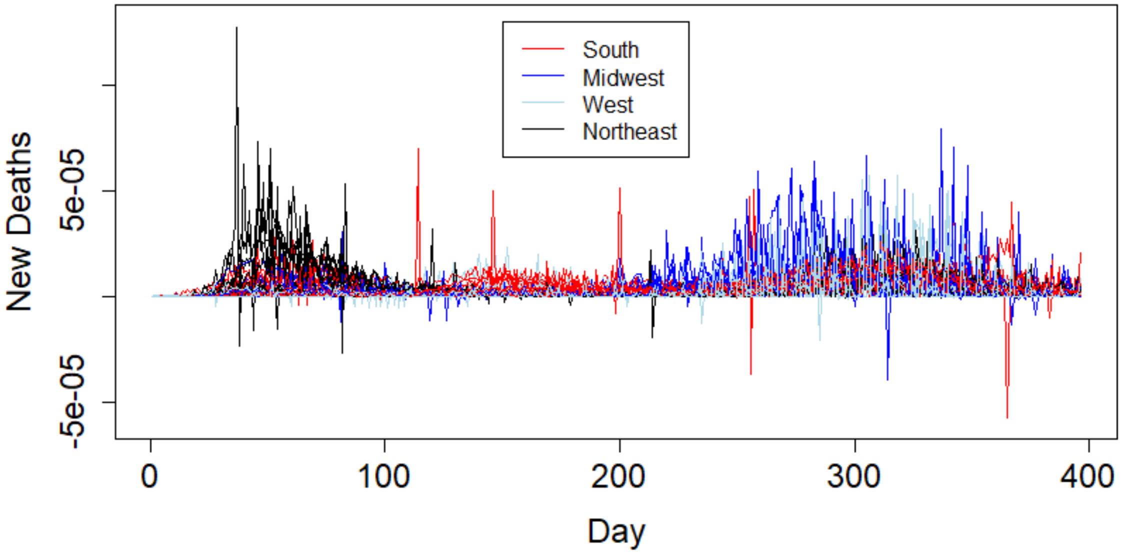

Figure 3.3.1-2, New COVID Deaths for States by Day

When looking at Figure 3.3.1-2, all of the lines on the graph seem to be following the same trends. There are some states that are higher and lower than others, but the new COVID deaths per state all seem to follow the same trend. In the graph, there are some states that stand out because the state has a lot higher or lower COVID deaths.

Figure 3.3.1-3, New COVID Deaths per Region by State by Day, displays the four regions and for each region, the states that are noticeable (have high or low COVID Deaths) in that region.

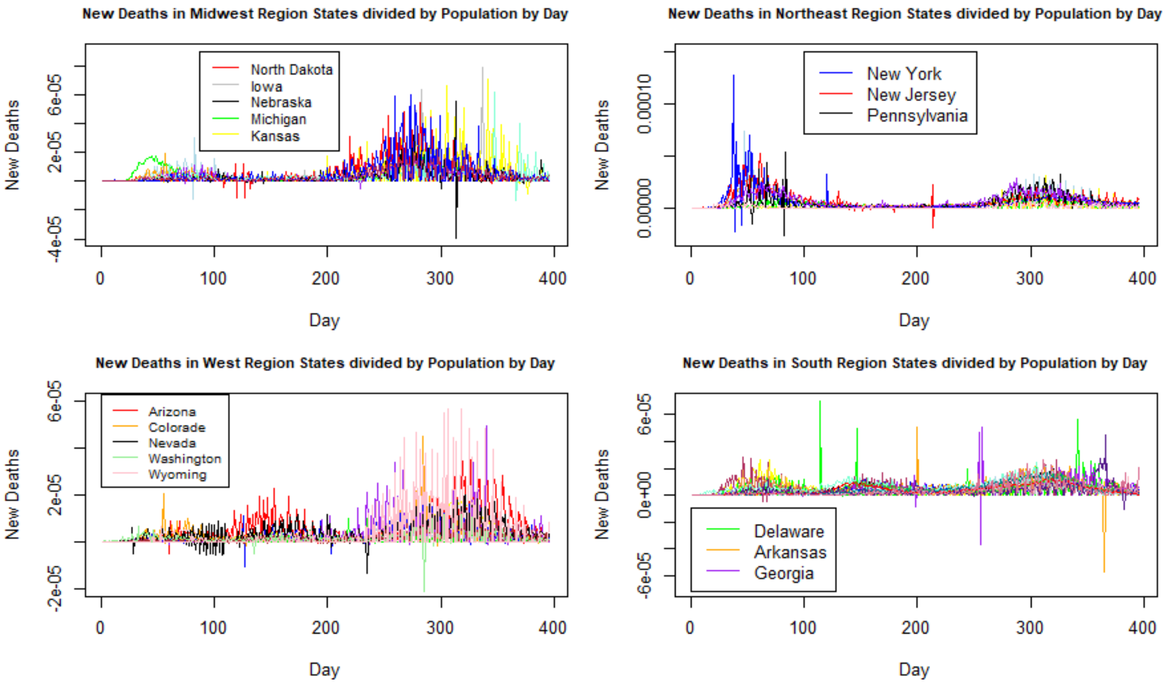

Figure 3.3.1-3, New COVID Deaths per Region by State by Day

Looking at the Midwest Region States first, North Dakota, Iowa, Nebraska, Michigan, and Kansas have spikes and the highest/lowest numbers of new COVID deaths. In the Northeast Region, New York, New Jersey, and Pennsylvania are the states that have some spikes or are the highest/lowest COVID deaths. In the West Region the states that stand out the most are Arizona, Colorado, Nevada, Washington, and Wyoming, around day 300 there are a lot of new COVID deaths. In the South region Delaware, Arkansas, and Georgia are the states that stand out the most. There are some spikes in the data for some of the states. Most of the states in this region also follow along the same pattern.

3.3.2 Smoothed Line New Deaths

This section displays the smoothed lines of the new COVID deaths data. In the beginning of section 3.1.2, it was analyzed and determined that the best smoothing lambda to use is 1e4. For this section and data, lambda = 10^4 will also be used to smooth the data.

Figure 3.3.2-1, Smoothed Line of New COVID Deaths per Region by Day, illustrates why a lambda of 4 is selected for the data.

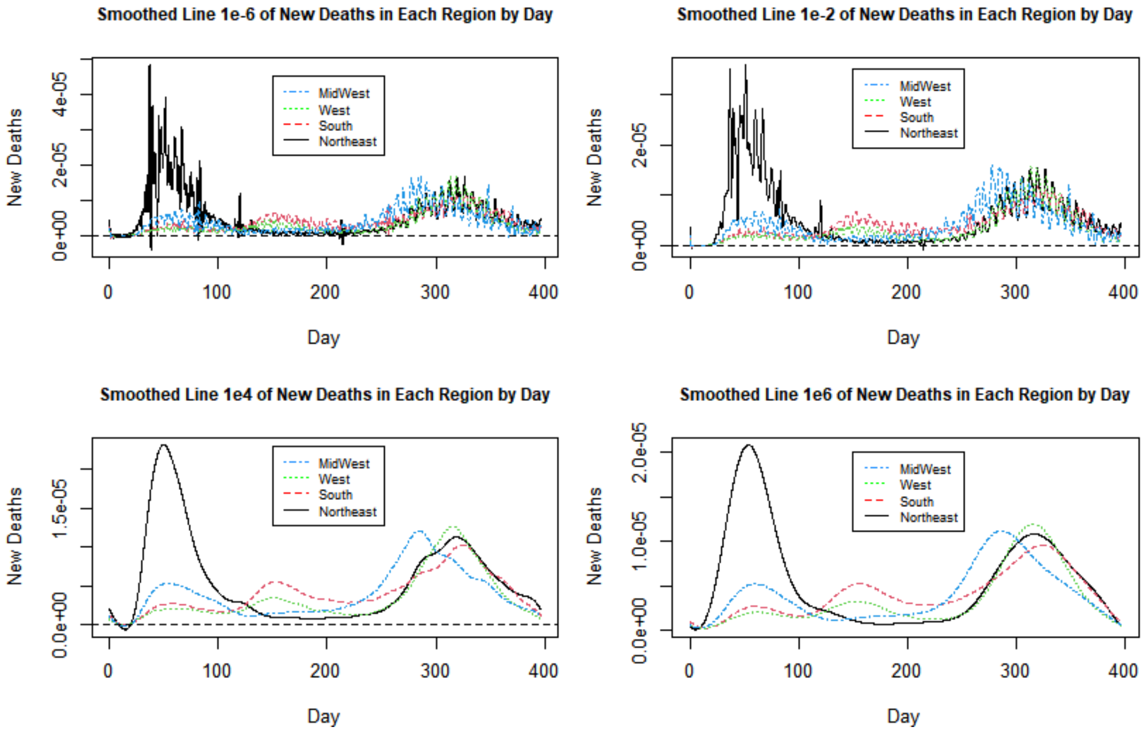

Figure 3.3.2-1, Smoothed Line of New COVID Deaths per Region by Day

In Figure 3.3.2-1 there are four different smoothed line graphs to evaluate and determine which one is the best fit for this data. The first smoothed line graph uses lambda = 10^{-6}. Here the graph shows lines still spike up and down with the data. The line needs to be more smoothed to be able to see the trend lines better. The second graph smooths the line out more, using lambda = 10^{-2}, but still has some up and down spikes and could look better. The next graph uses lambda = 10^{4}, this graph looks a lot better with the data and lines smoothed out more. The data is smoothed out one more time using lambda = 10^{6}. For this lambda value, the graph is too smooth. The data needs to have some curves, to be able to see what is going on and to be able to analyze the data better. When looking at Figure 3.1.2-1 and Figure 3.3.2-1 the best lambda that will be utilized in this data analysis is 10^{4}.

Figure 3.3.2-2, Smoothed Line of New COVID Deaths per Region by Day, displays the smoothed lines of the new COVID Deaths per region.

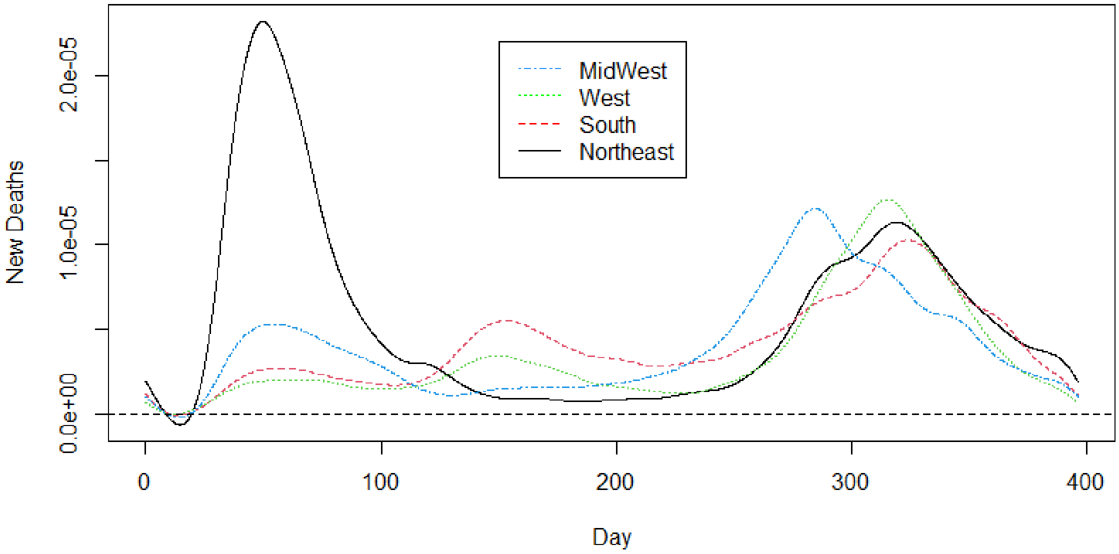

Figure 3.3.2-2, Smoothed Line of New COVID Deaths per Region by Day

Figure 3.3.2-2, the Northeast area spikes up right away and then goes back down. Around day 150 the South region goes up higher than the other regions do. Then around day 250 all regions start to spike up with all regions having about the same number of new COVID deaths.

In Figure 3.3.2-3, Smoothed Line of New COVID Deaths for States by Day, the deaths in the states are being displayed.

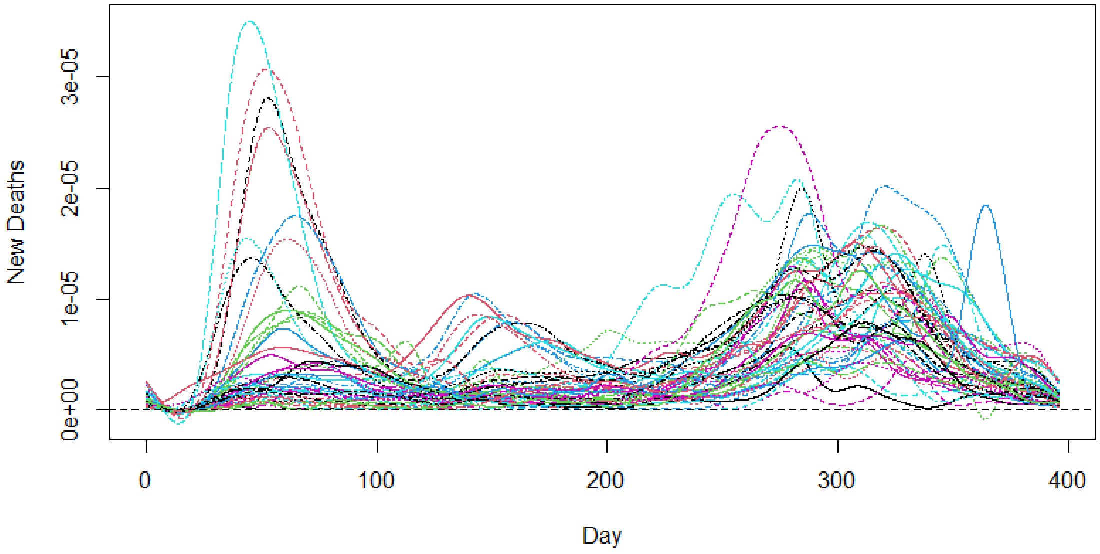

Figure 3.3.2-3, Smoothed Line of New COVID Deaths for States by Day

In Figure 3.3.2-3, a few states stand out more compared to the others. There are three different spikes in the data around day 50, day 150, and day 250-350. The states follow along the same path with some states standing out from the rest.

3.3.3 Smoothed Line Graph of New Deaths

This section looks at the new COVID deaths over time and the average over time for all the data. Figure 3.3.3-1, Smoothed Line of New COVID Deaths with the Mean Line per Region by Day, shows the number of new COVID deaths in that region for that day over time.

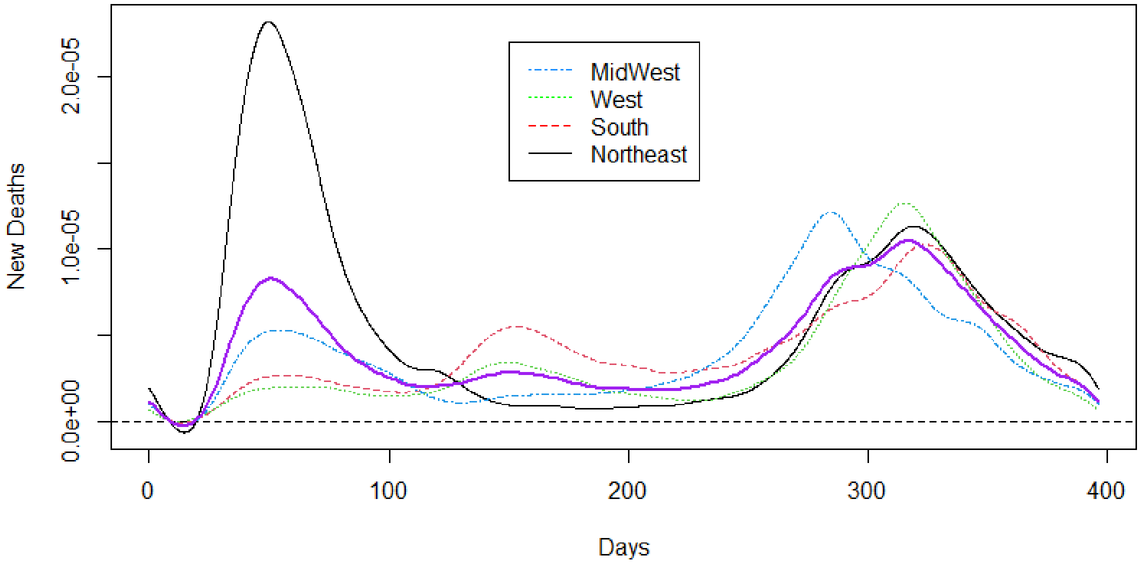

Figure 3.3.3-1, Smoothed Line of New COVID Deaths with the Mean Line per Region by Day

The purple line shows the mean positive cases for all the regions over time for that day. The average line appears to be in the middle of the four regions, following the same trend as the other data. Around day 50 the Northeast region pulls the average up to make it higher than the other regions. Around day 150 the South is up a lot higher than the other regions.

Figure 3.3.3-2, Smoothed Line of New COVID Deaths with the Mean Line for States by Day, show the smoothed lines of the new COVID deaths. On this graph, the mean line of the data is plotted in a solid bold purple line as well as the new deaths by states lines. The mean line of the data helps to analyze where the average of the data is and see which states stand out from the average.

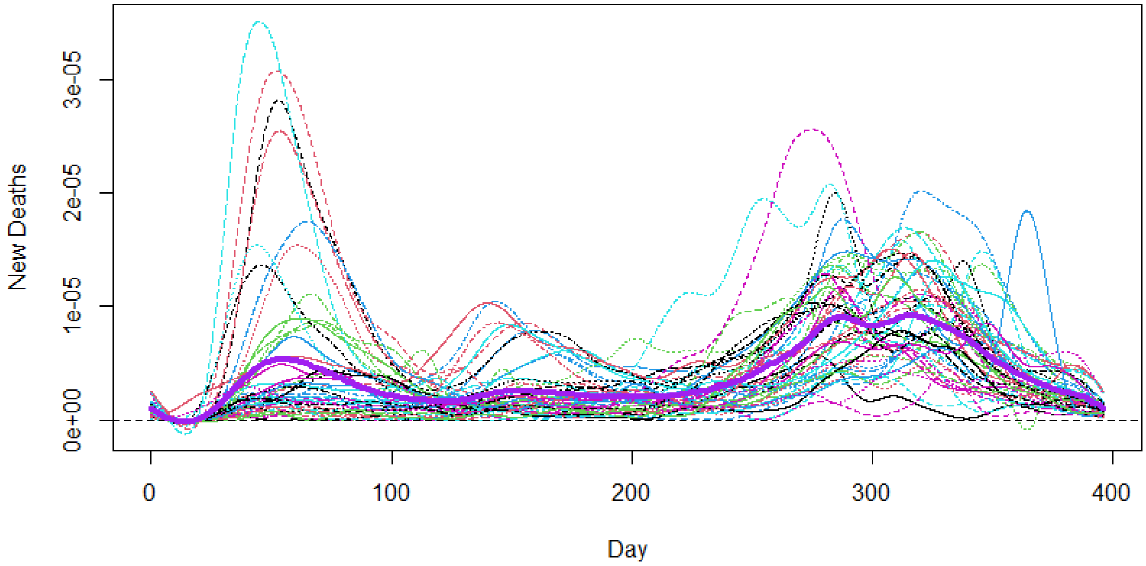

Figure 3.3.3-2, Smoothed Line of New COVID Deaths with the Mean Line for States by Day

Figure 3.3.3-2 shows that around days 50, 110, and 250 to 350 there are spikes in the data. For those days the mean line goes up but not as high as some of the data does. There are some states that pull the mean line up, compared to the other states that have a lower COVID death count.

The next part of the data is the smoothed lines of the data minus the mean line. By plotting the results, Figure 3.3.3-3 displays which regions are above or below the average of the data.

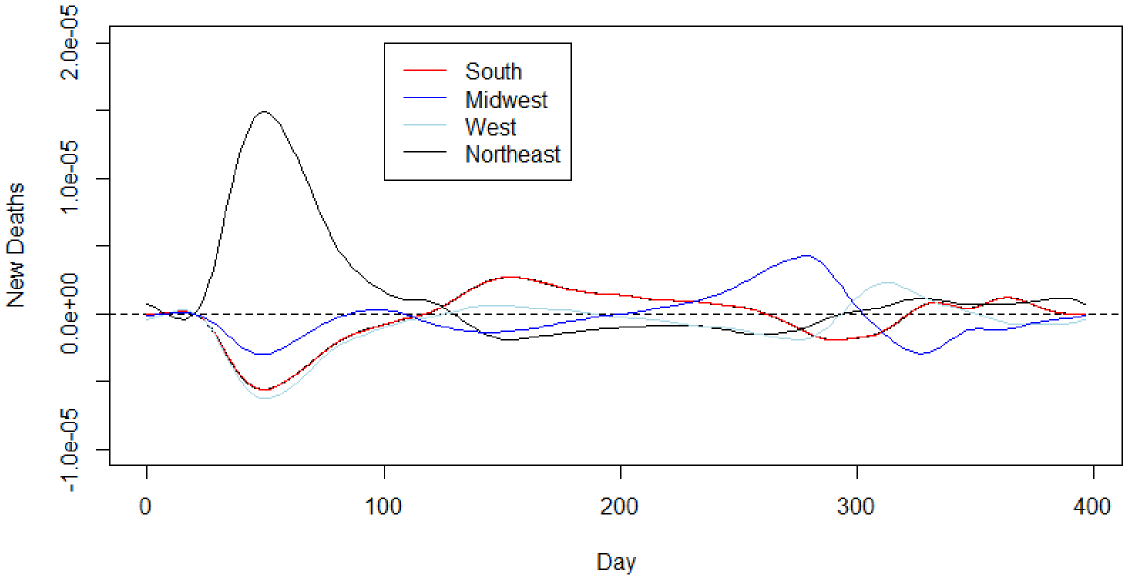

Figure 3.3.3-3, Smoothed Line of New COVID Deaths minus the Mean per Region by Day

Figure 3.3.3-3 shows that the Northeast area is again the highest in the beginning as is shown in Figure 3.3.3-1. The South region is higher around day 150, then drops down and comes back up at the end. The West is low at first and then stays constant going up and down around 0. The Midwest stands out the most around day 280 that spikes up higher than the rest of the regions.

The last Figure 3.3.3-4, Smoothed Line of New COVID Deaths minus the Mean for States by Day, shows the number of cases in each region by the states minus the averages.

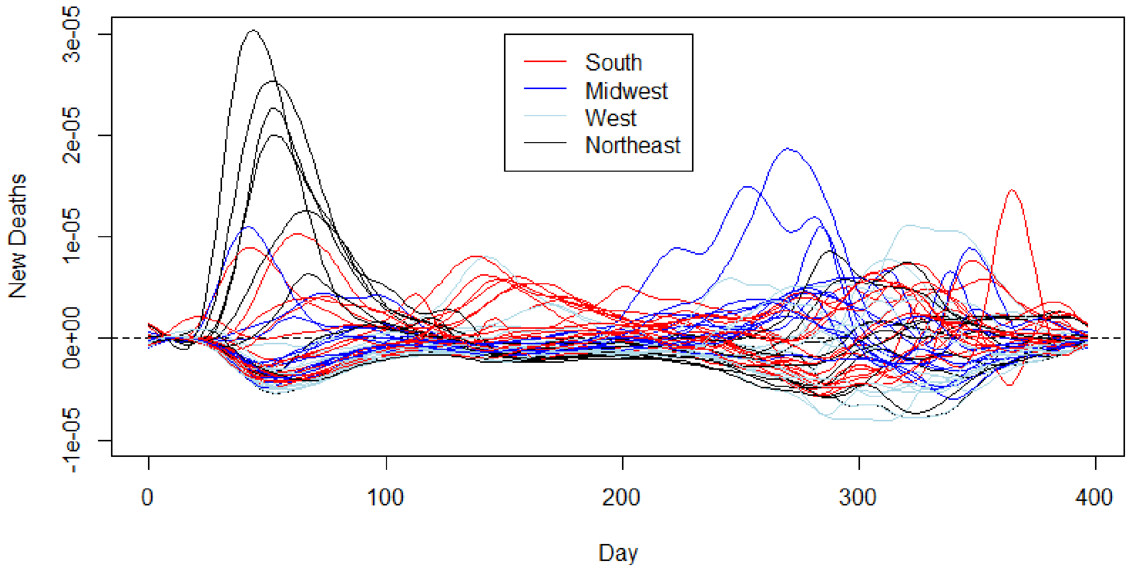

Figure 3.3.3-4, Smoothed Line of New COVID Deaths minus the Mean for States by Day

Looking at the last Figure 3.3.3-4, in the beginning there are a few Northeast and South states that are above the average that stand out. Around day 150 it switches, and a West state and South states stand out above the average and the other states. Then around day 250 the Midwest states are the highest and really stand out above the rest. From there as time goes on the states flatten out to be about the same around the 0 and the average.

3.3.4 Variance-Covariance of New Deaths

In Figure 3.3.4-1, Contour Plot of New COVID Deaths per Region by Day shows the contour plot of the variance-covariance of the new COVID death cases per region by day.

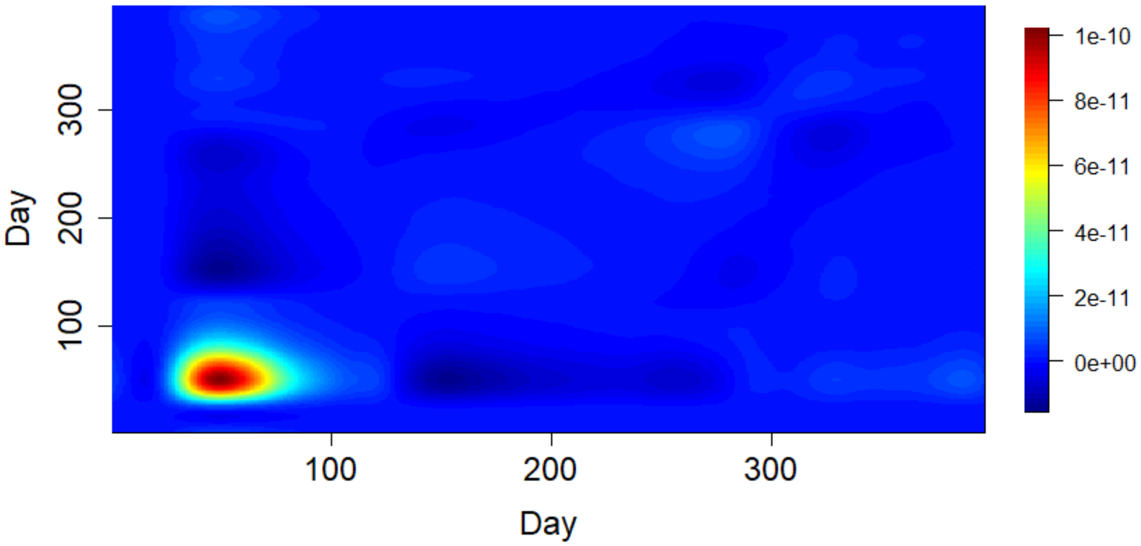

Figure 3.3.4-1, Contour Plot of New COVID Deaths per Region by Day

This plot shows the covariance surface of the new COVID deaths per region by each day. The diagonal of the plot is also the variance of the data. On the diagonal of the plot, for each day the variance counts are a mirror image of each other. The dark red area on the plot is the highest hospitalized patients which is 1e-10, and is along the diagonal which is part of the variance.

Figure 3.3.4-2, Contour Plot of New COVID Deaths for States by Day, displays a contour plot of the variance-covariance of the new COVID deaths for states by day.

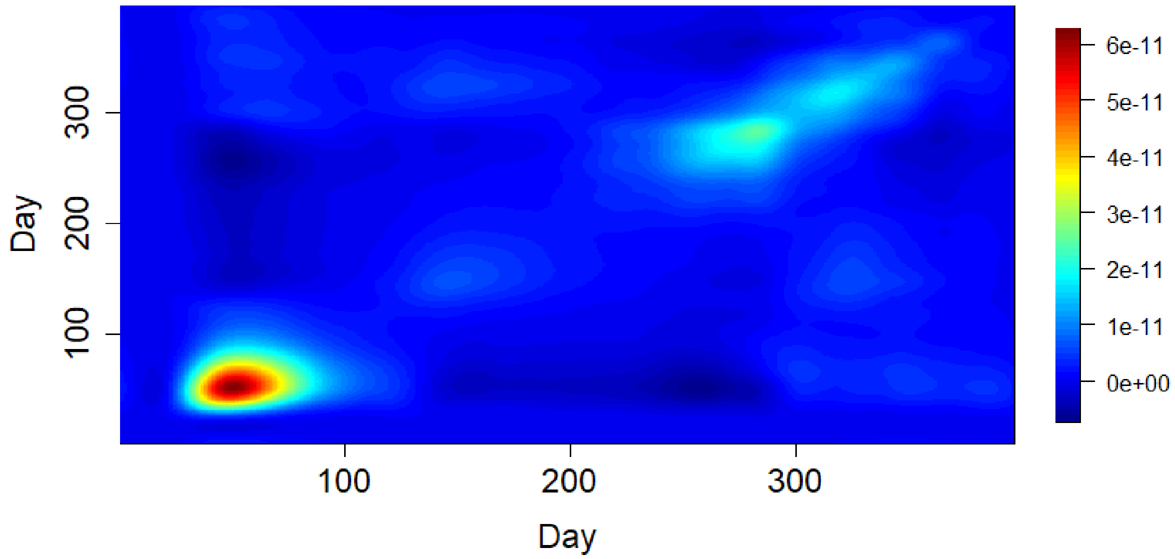

Figure 3.3.4-2, Contour Plot of New COVID Deaths for States by Day

This plot shows the covariance surface of the new COVID deaths for states by each day. The diagonal of the plot is the variance of the data. Across the diagonal of the plot on each side they mimic each other. The dark red of the plot is the highest count which is 6e-11, this part is along the diagonal which is part of the variance. The next section will look at the correlation coefficient plots which are easier to read and to tell more from about the data.

3.3.5 Correlation Coefficient Plot New Deaths

This section investigates the correlation coefficient plots of the new COVID deaths over time. The X and the Y axis are both days to compare the time of year and the number of new COVID deaths there have been. Figure 3.3.5-1, Correlation Coefficient Plot of New COVID Deaths per Region by Day, displays the correlation coefficient plot of the new COVID deaths per region by day.

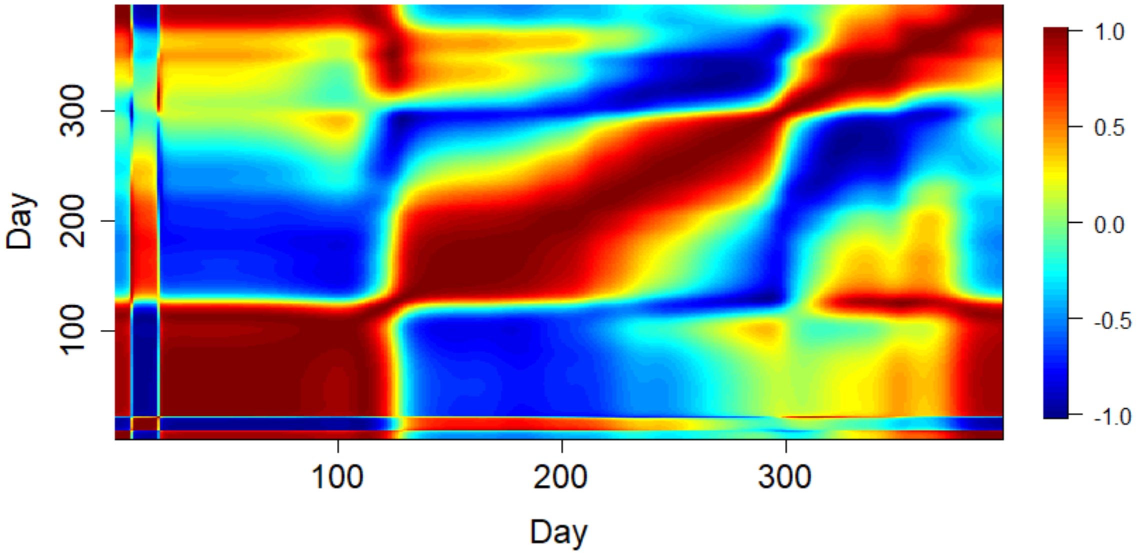

Figure 3.3.5-1, Correlation Coefficient Plot of New COVID Deaths per Region by Day

Figure 3.3.5-1 is easier to read and conveys more than the variance-covariance plot that is in Figure 3.3.4-1, Contour Plot of New COVID Deaths per Region by Day, in the previous section. Figure 3.3.5-1, along the diagonal, is equal to the value 1 which is the maximum correlation. The highest value is along the diagonal when it is 1, the lowest correlation is the darker blue (-1.0) which is the correlation between June to August (x-axis) and April to June (y-axis) in the region areas. There is also a small spot of dark blue which is the correlation between December to January (x-axis) to November (y-axis)

Figure 3.3.5-2, Correlation Coefficient Plot of New COVID Deaths for States by Day, is the correlation coefficient plot of the new COVID deaths for states by day. Along the diagonal is again, the highest correlation value of 1. The smallest correlation is in the dark blue again of -0.5, which is the correlation from April to June (x-axis) to March (y-axis) in the states areas.

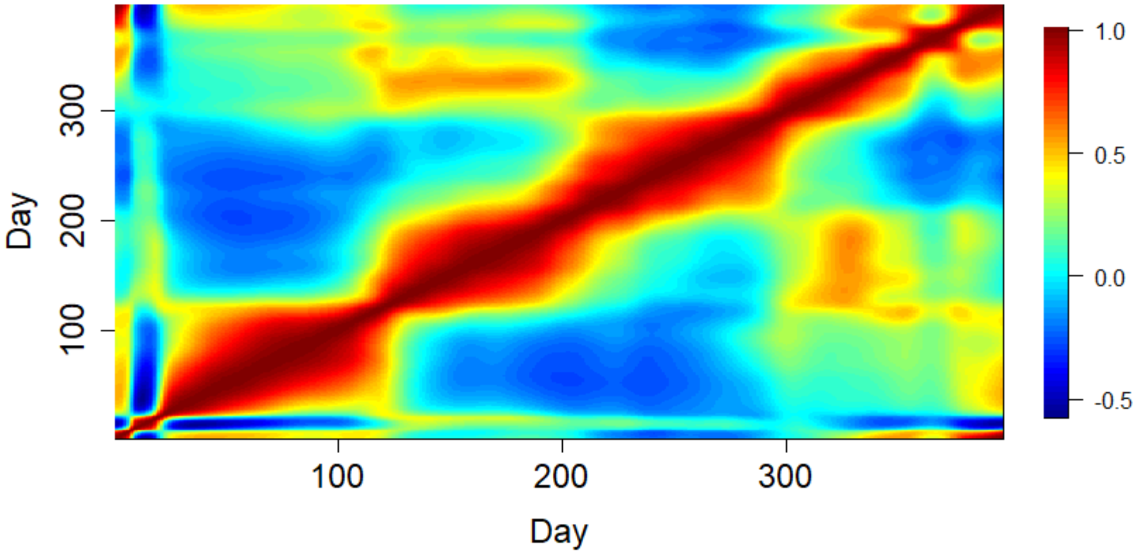

Figure 3.3.5-2, Correlation Coefficient Plot of New COVID Deaths for States by Day

Figure 3.3.5-2 has brighter colors (green, yellow, orange, and red) than the darker blue colors. This shows that there is more of a correlation in the new COVID deaths over time than what there is in the new COVID cases over time (Figure 3.1.5-2).

3.3.6 Eigenvalue Plot New Deaths

This section looks at the Functional Principal Component Analysis (FPCA) and the eigenvalues of the new COVID deaths. All the data sets use 4 Functional Principal Components (FPC) which was discussed in section 3.1.8. Figure 3.3.6-1, Eigenvalue Plot of New COVID Deaths per Region by Day, looks at each eigenvalue for each FPCA.

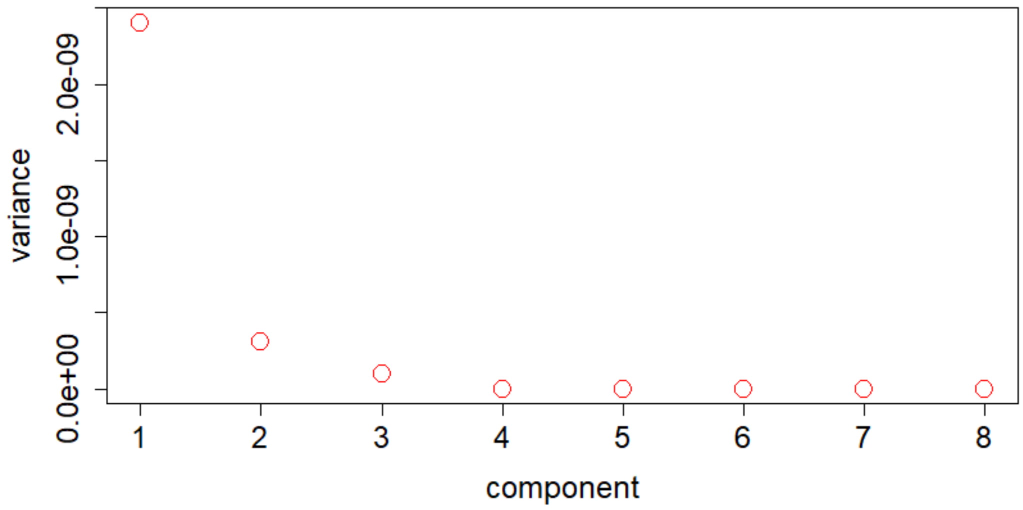

Figure 3.3.6-1, Eigenvalue Plot of New COVID Deaths per Region by Day

Each eigenvalue is equal to the variance of the FPC score. As the eigenvalue component increases the variance value decreases and gets closer and closer to 0. The first eigenvalue is large and then the next one drops in value very fast. Figure 3.3.6-2, Eigenvalue Plot of New COVID Deaths for States by Day, look at the same values but for the new COVID deaths for states by day.

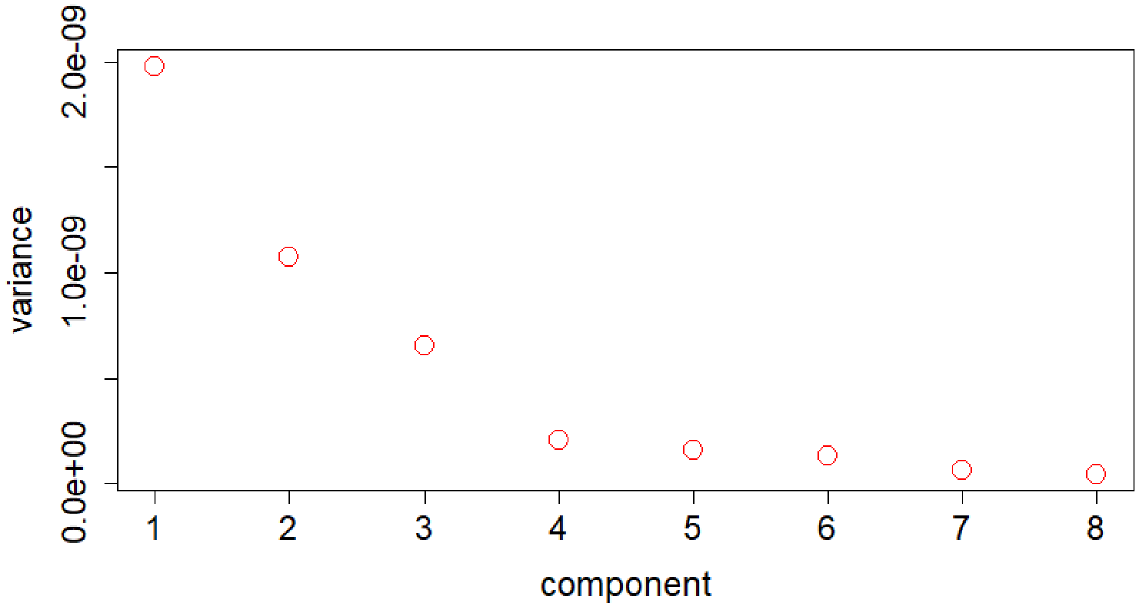

Figure 3.3.6-2, Eigenvalue Plot of New COVID Deaths for States by Day

The values and eigenvalues for the new COVID deaths for states by day are bigger values than the ones for new COVID deaths per region. The eigenvalues for Figure 3.3.6-2 drop a little slower than in Figure 3.3.6-1. The eigenvalues don’t seem to get to around 0 until the 7th eigenvalue.

3.3.7 Cumulative Percentage Plot Deaths

This section looks at the cumulative percentage explained for the total variation plots of the new COVID deaths. Theses plots are the sum of the ten eigenvalues then divided by the sum of all the eigenvalues. This shows what number of FPC's (x-axis) are needed to show what percentage of the data (y-axis).

Figure 3.3.7-1, Cumulative Percentage Plot of New COVID Deaths per Region by Day, displays the cumulative percentage of the new COVID deaths per region by day.

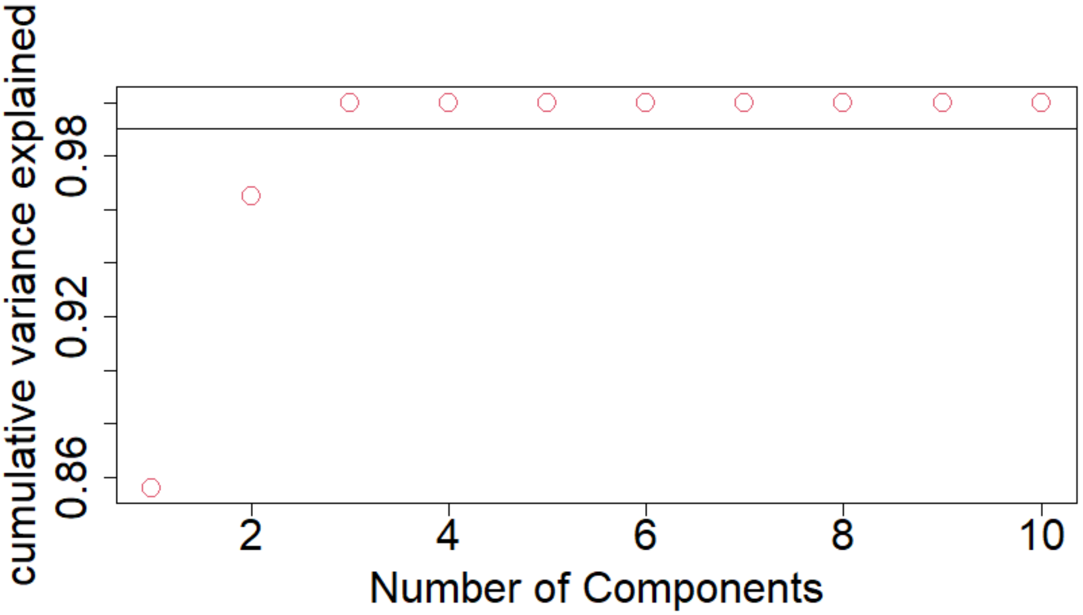

Figure 3.3.7-1, Cumulative Percentage Plot of New COVID Deaths per Region by Day

Figure 3.3.7-1 shows that the first 3 FPCS explain about 99% of the total variation of the data. Once the graph reaches the 3rd and above FPCS the graph shows that the value is above the 99% line and remains about the same value. With this data the first 3 FPCs are needed to explain most of the data.

Figure 3.3.7-2, Cumulative Percentage Plot of New COVID Deaths for States by Day, looks at the cumulative percentage of the new COVID deaths for states by day.

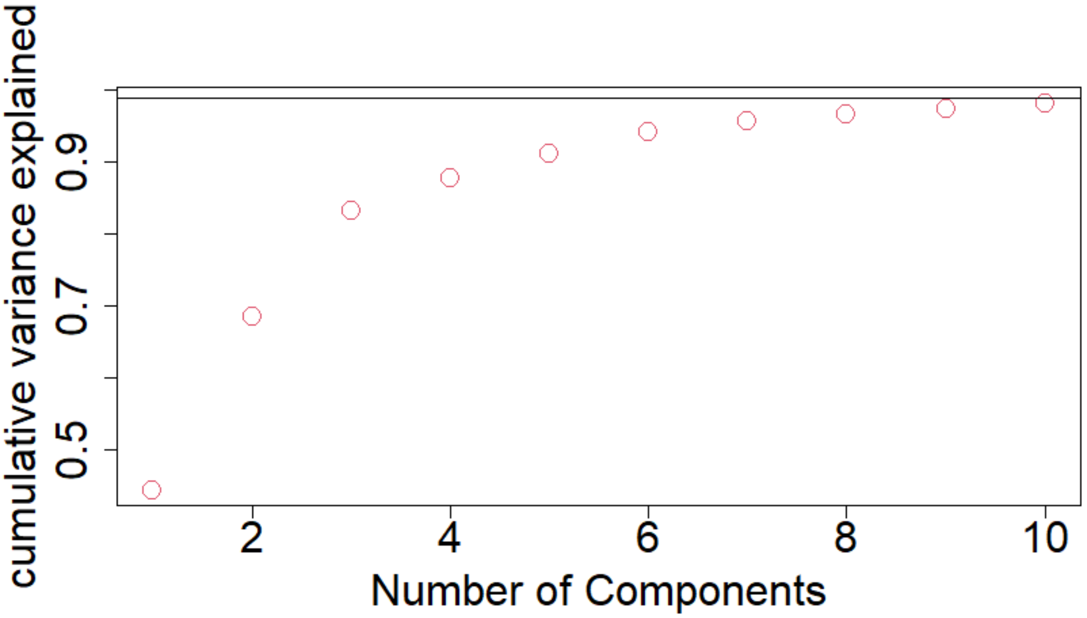

Figure 3.3.7-2, Cumulative Percentage Plot of New COVID Deaths for States by Day

Figure 3.3.7-2 displays that more FPCs are needed to explain the data than in Figure 3.3.7-1. For the new COVID deaths for states by day, the number of components does not go above the 1 line until about 10 components. This data would need about 10 FPCs to explain about 99% of the data.

3.3.8 Functional Principal Component Functions Graph of Deaths

Section 3.3.7 presented cumulative percentages and how many FPCs are needed to explain the data. Looking at Figures 3.3.7-1 it shows that 3 are needed and 3.3.7-2 shows that 10 are needed. As discussed in section 3.1.8, 4 FPCs will be used to stay consistent and to look further into the data.

Figure 3.3.8-1, Functional Principal Component Functions of New COVID Deaths per Region by Day, looks deeper into explaining the data.

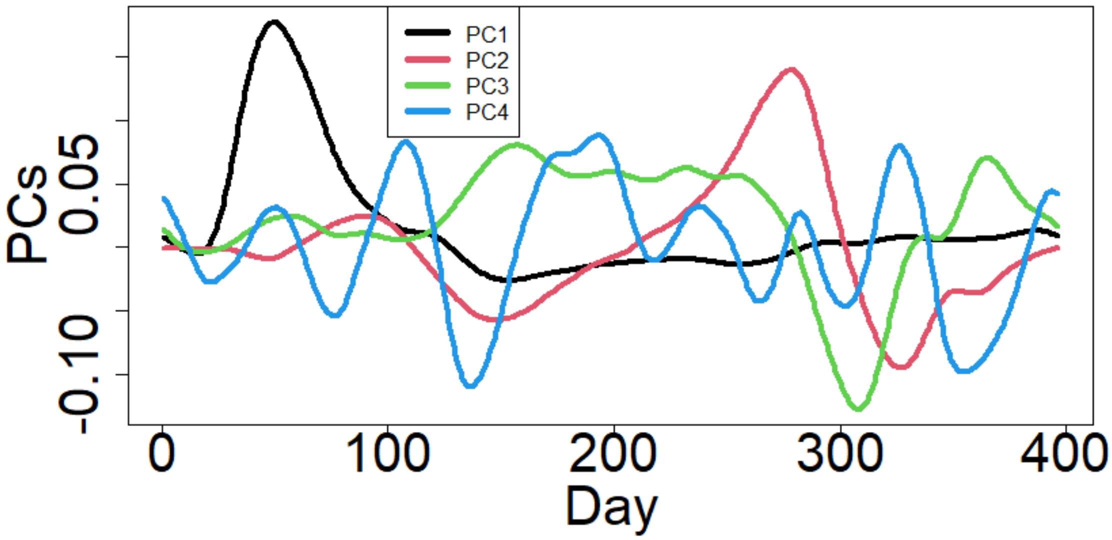

Figure 3.3.8-1, Functional Principal Component Functions of New COVID Deaths per Region by Day

In Figure 3.3.8-1, PC1 matches up with the Northeast region, they both spike up the highest right off the bat. PC2, the red line, matches up the best with the Midwest region, staying down low most of the time until it spikes up around day 300. PC3, the green line, matches up best with the south region, continuously going up and down and never really staying constant. The last PC4 then fits best with the west region, spiking up around day 150 and then staying lower than the other regions.

Figure 3.3.8-2, Functional Principal Component Functions of New COVID Deaths for States by Day, displays the four Principal Components that best represents/explains the data.

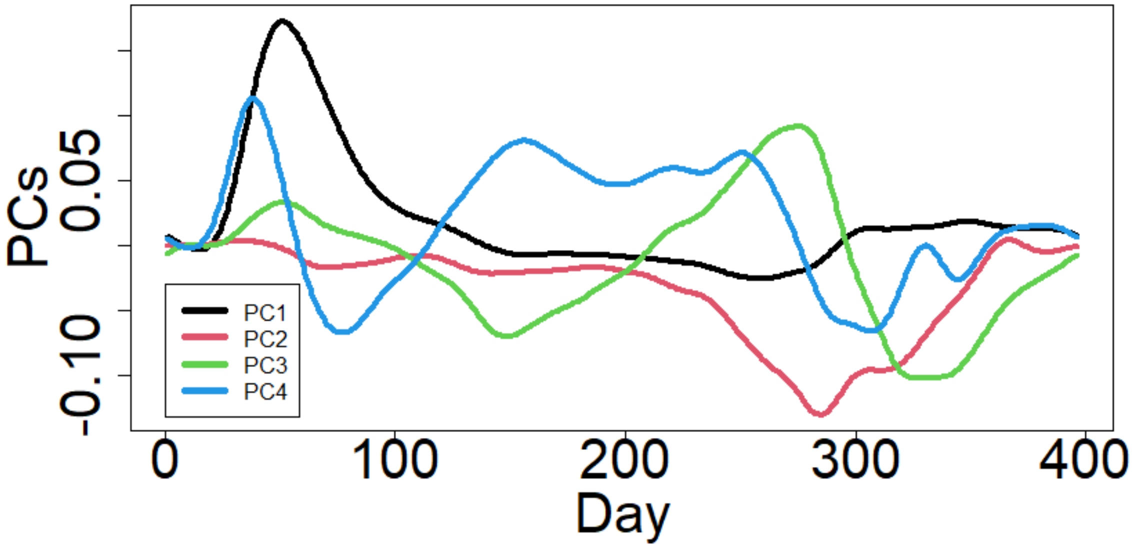

Figure 3.3.8-2, Functional Principal Component Functions of New COVID Deaths for States by Day

Figure 3.3.8-2 is similar to Figure 3.3.8-1 (which represents Region Data), where Figure 3.3.8-2 represents data for each state and not just the regions. Referencing Figure 3.3.1-3, New COVID Deaths per Region by State by Day, some of the states that stand out match up with some of the Principal Components (PC). Michigan and Colorado matches up best with PC1, Kansas and Washington with PC2, Georgia and Wyoming with PC3, and Arizona and New York with PC4. All of the states will match up with one of the PC’s but those examples just show some of the states that stood out from the others.

3.3.9 FPC1 vs FPC2 Graphs of Deaths

This section investigates the FPC1 on the x-axis and the FPC2 on the y-axis. This shows what states and regions are grouped together and are similar and which ones stand out from the others. Figure 3.3.9-1, FPC1 vs FPC2 Scores of New COVID Deaths per Region by Day, displays FPC1 (x-axis) against FPC2 (y-axis).

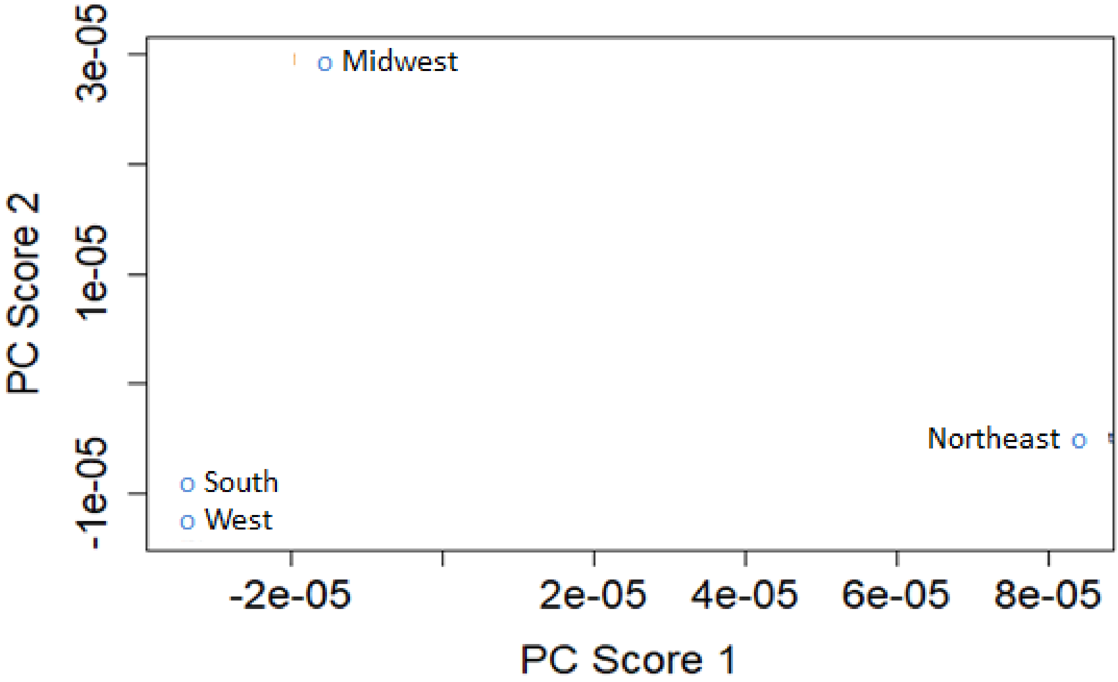

Figure 3.3.9-1, FPC1 vs FPC2 Scores of New COVID Deaths per Region by Day

Looking at Figure 3.3.9-1 one can see that the South and the West regions are grouped together. With those regions grouped together, it shows that those two regions are similar with new COVID deaths. The Northeast and the Midwest are not grouped with any other region, showing that these regions aren't like the other regions in the new COVID deaths. Figure 3.3.9-2 looks at the same type of graph but with the different states. This shows what states are similar in the new COVID deaths. It also shows which states are outliers and aren't similar to the other states.

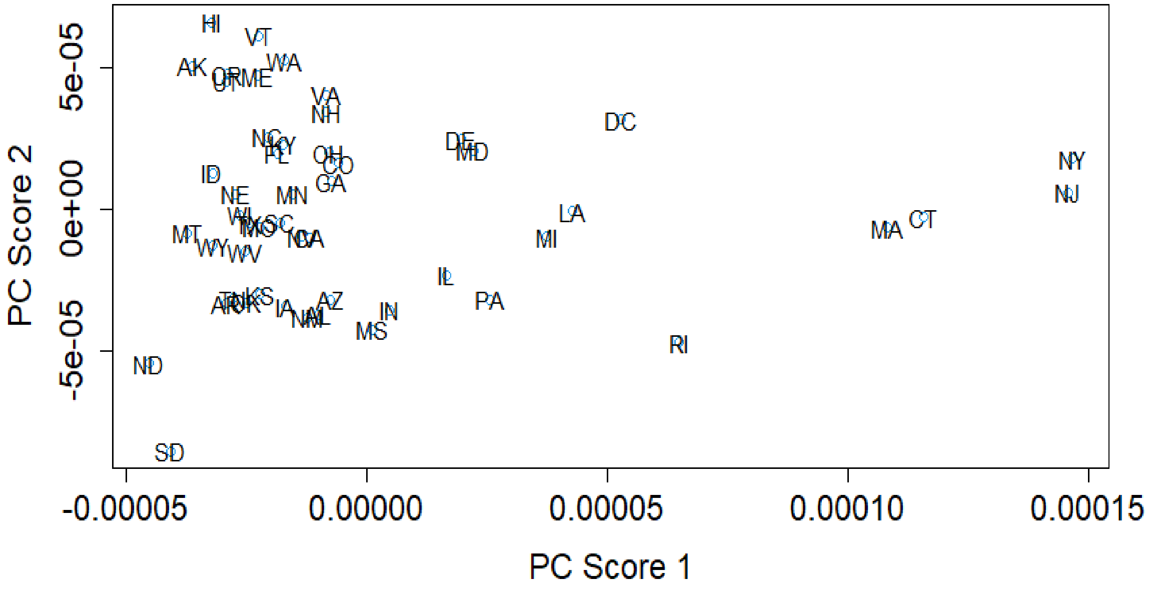

Figure 3.3.9-2, FPC1 vs FPC2 Scores of New COVID Deaths for States by Day

Figure 3.3.9-2, most of the states are grouped together but there are a few that stand out from the others. Massachusetts, Connecticut, New Jersey and New York are off to the top right from the other states, this shows that those states didn’t follow along with the other states in the new COVID deaths. New Jersey and New York are close together and are the same with the new COVID deaths. Massachusetts and Connecticut are close together and are similar in the new COVID deaths. South Dakota is one state that is a little lower than the other states and stands out a little more. Rhode Island is also a little further out and away from the group of states.

3.3.10 Regression Coefficients

Figure 3.3.10-1, New COVID Deaths Mean, Difference and Predicted by Region by Day, displays the regression coefficients and the differences between the mean and each region new COVID deaths.

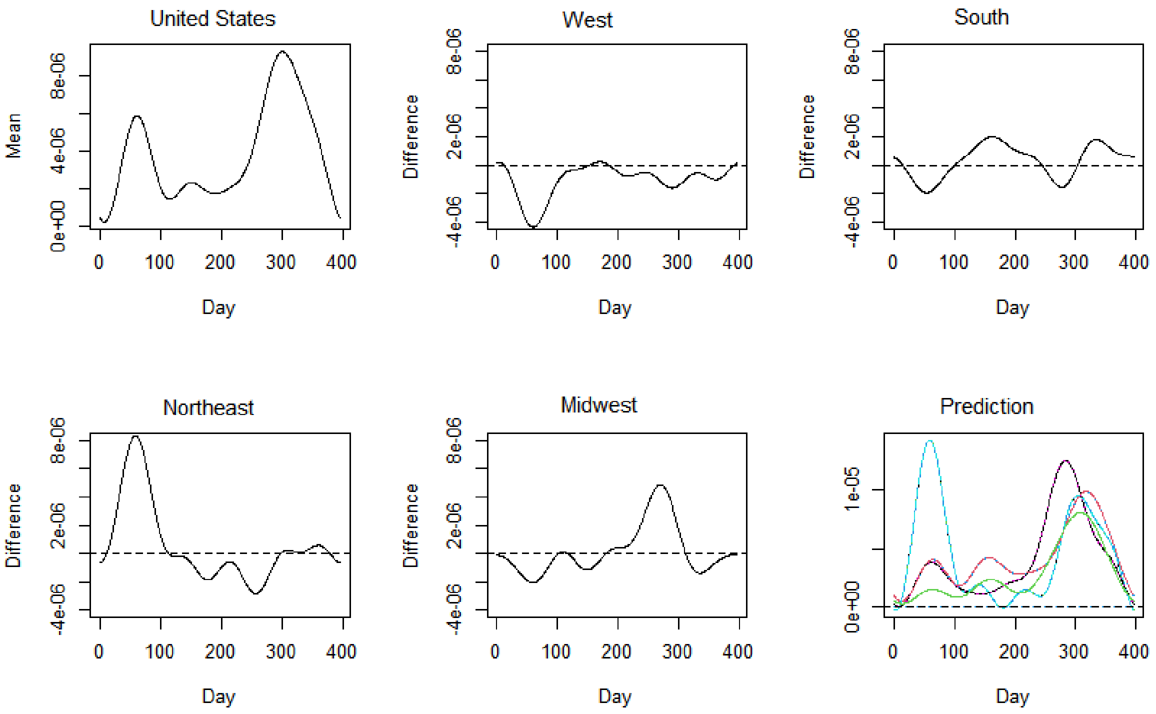

Figure 3.3.10-1, New COVID Deaths Mean, Difference and Predicted by Region by Day

The United States graph in Figure 3.3.10-1 shows the mean of all the new COVID deaths over the time frame. The next four graphs show the difference in that area of the new COVID deaths and the mean from the United States. With the four regions, when they are in the positive values it shows when they have the highest impacts. When the values are in the positive it shows that they had more COVID deaths in that region and what the mean is for that time. The Northeast region stands out in the first 100 days being higher than the average. The South region stands out in the middle between days 100 to 250 and 300 to 396. The Midwest region stands out in the middle between 290 and 300 days. The West region most of the time stays below the 0 line and doesn't have that much impact on deaths. The last graph shows the predictions or the new COVID deaths for that region.

3.3.11 Functional F-Test

This last section looks at the Functional F-Test for the new COVID deaths over time. As was noted earlier, for this data, consider the hypotheses:

- Null Hypothesis: H_0: There is no difference in the regions for new COVID deaths
- Alternative Hypothesis: H_a: There is a difference in the regions for new COVID deaths

Figure 3.3.11-1, New COVID Deaths for Functional Permutation F-Test Statistic by Day, shows if there is a difference between the regions with the new COVID deaths.

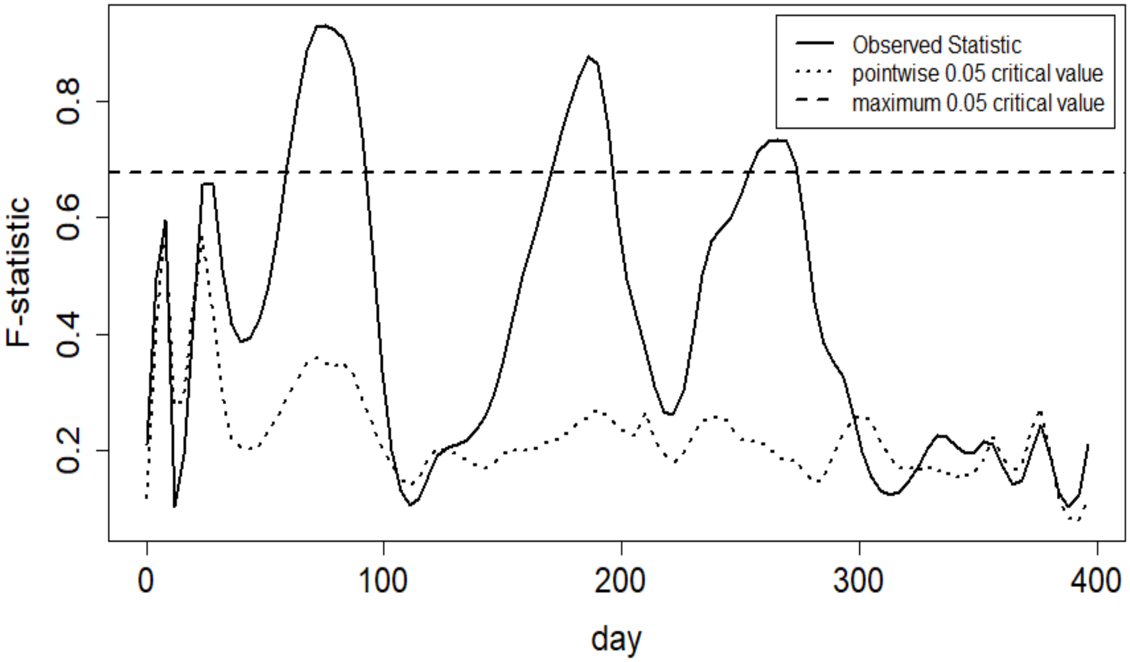

Figure 3.3.11-1, New COVID Deaths for Functional Permutation F-Test Statistic by Day

In this figure it shows the solid line as the observed statistic of the new COVID deaths. If the observed statistic is below the dotted pointwise 0.05 critical value, then there is no regional effect or difference in that time period. The figure shows that the observed statistic is above the pointwise critical line which means that there is a difference in regions for the new COVID deaths. From day 0 to 30, 100 to 120, 300 to 320, and 360 to 380 are the times that the observed statistics line falls below the pointwise value that there isn’t a difference in new COVID deaths for the different regions. For most of this data the F statistic (Observed Statistic) is above the pointwise 0.05 critical value, this means that we can reject the null hypothesis and conclude that there is a difference in the regions for new COVID deaths.

Figure 3.3.11-2, New COVID Deaths for Functional 5,000 Permutations F-Test Statistic by Day with Bonferroni Correction, shows if there is a difference between the regions with the new COVID deaths adjusting for the different sets of data.

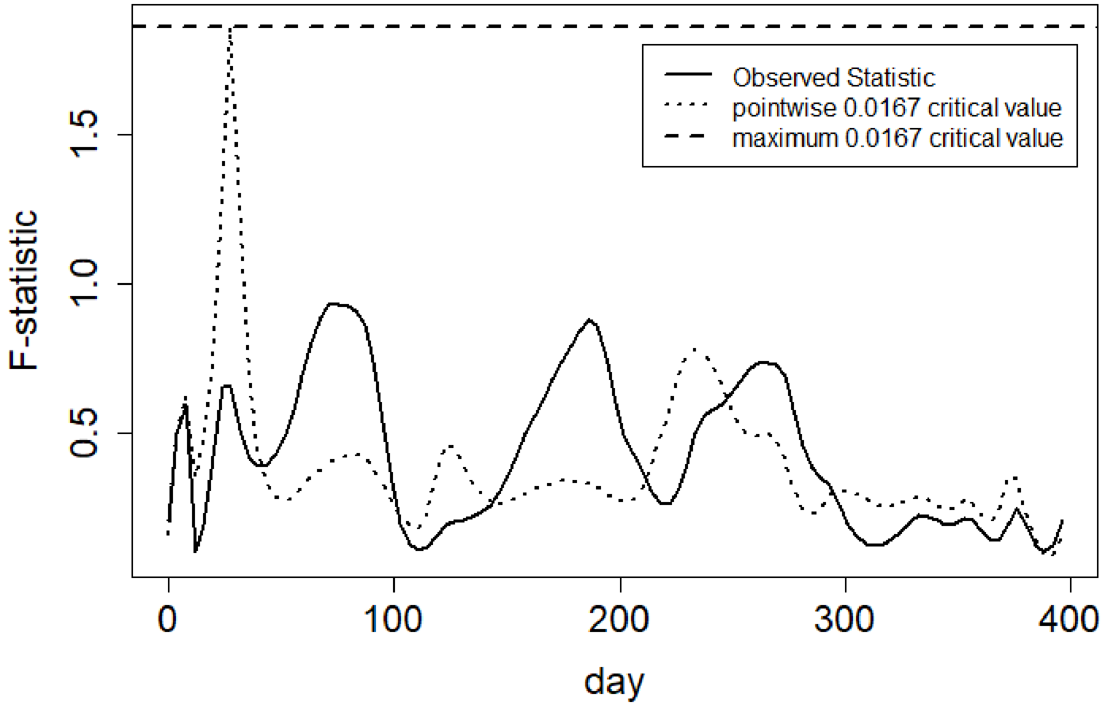

Figure 3.3.11-2, New COVID Deaths for Functional 5,000 Permutation F-Test Statistic by Day with Bonferroni Correction

In this figure with the Bonferroni correction, it shows for all the data over time there is a difference in the regions for new COVID deaths. In Figure 3.3.11-1, it shows around day 0 to 30, 100 to 120, 300 to 320, and 360 to 380 the data is below the pointwise 0.05 critical value. With the Bonferroni correction of the different data sets, the observed line is still below the 0.0167 critical line. With the Bonferroni correction in the F-test the data is a lot more sensitive now. With the data being more sensitive the observed line falls below the pointwise line more often than it did without the Bonferroni adjustment. Figure 3.3.11-2 shows the data with 5,000 permutations of the data. A permutation is putting the data in different arrangements of the data to get a more accurate output of the F-Test.

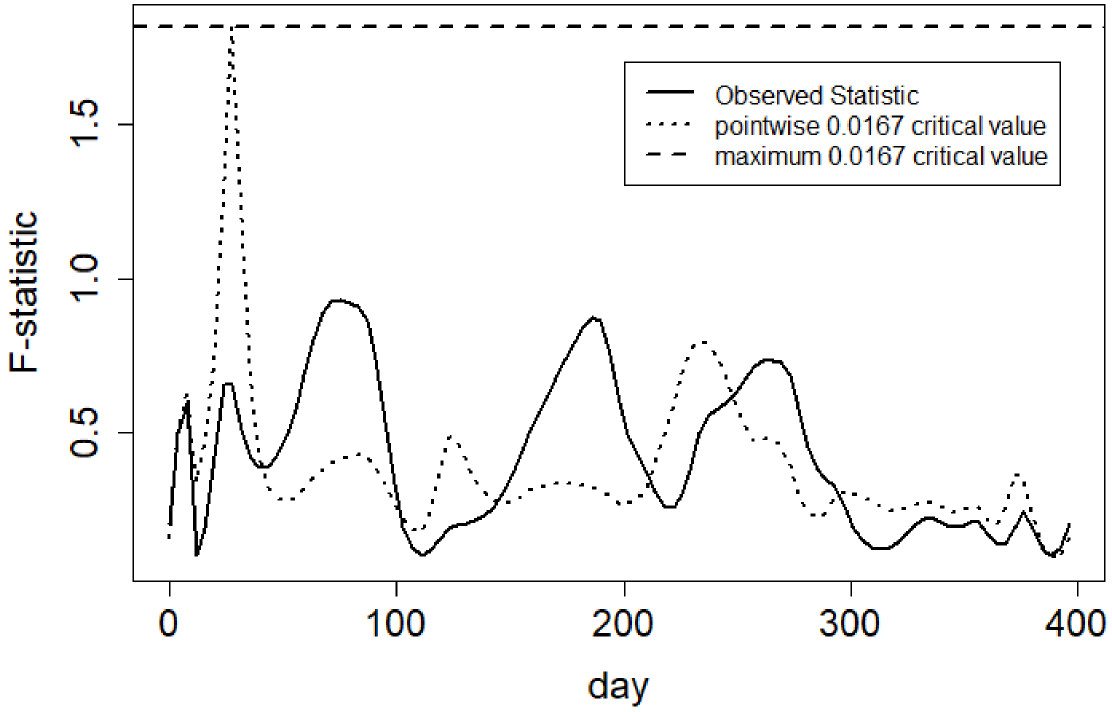

Figure 3.3.11-3, New COVID Deaths for Functional 10,000 Permutation F-Test Statistic by Day with Bonferroni Correction

Figure 3.3.11-3 shows the data with 10,000 permutations to try to get the F-Test value to be more accurate. Comparing Figure 3.3.11-2 and Figure 3.3.11-3, the outputs for both look to be about the same. Since 5,000 and 10,000 permutations are the same this shows that using 5,000 permutations is accurate enough to use.

4.0 Discussion

Analyzing the different sets of data:

- new COVID cases,
- hospitalized COVID patients, and
- new COVID deaths.

Each data set appears to show differences for the four different regions and states for the new COVID cases, hospitalized COVID patients, and new COVID deaths. In some of the data sets, there is a larger difference in the states and the regions than in other data sets. Some states and regions are more similar to each other than other states. Analyzing the FPCA Figures for the different data sets, the South and West regions are always together. With them always being close together they are similar in their outcomes. The Northeast and Midwest were always off by themselves and not like other regions. When looking at the FPCA Figures for the states, depending on the data type--new COVID cases, hospitalized patients, and new COVID deaths--different states stand out for each one. They aren't always the same states that stand out when looking at the different data types. This could be happening from the way that the different regions are handling COVID by wearing a mask. The Northeast and the Midwest regions might always be together and similar in the new COVID cases, hospitalized patients, and new COIVD deaths from the different ways that they are handling COVID. Some of the states that are close to each other might be similar since they might be handling COVID the same or it might be carrying over from one state to another. For example, if one of the states has high COVID cases, it might be causing the states next to them to be increasing in the number of COVID cases they have as well. Depending on the time of year there are more new cases, hospitalized patients, and deaths than at other times of the year. During the holiday seasons (November to January) the new cases, hospitalized patients, and deaths are a lot higher than in the other months.

After seeing the outcomes from the different types of data and the region/state data, more analysis could be done to find out more. The data could be analyzed down to the county level to be more granular with the different states and the data. The data could also be looked at from a political view, looking at the different states to see their political party and to see from those states how similar or different they are from each other. Another way the data could be analyzed is by looking at the different regions/states and clustering the data together to see how similar or different the states are. The data could also have been collected longer when the vaccines started to come out. With the vaccines, the data then can be looked at to see if the new cases, hospitalized patients, and new deaths got better or worse in consideration.

References

1. “Basics of Covid-19.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 24 May 2021, www.cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19/basics-covid-19.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fcdcresponse%2Fabout-COVID-19.html
2. Bureau, US Census. “Idaho Is NATION'S Fastest-Growing State, Census Bureau Reports.” *The United States Census Bureau*, 29 Mar. 2018, www.census.gov/newsroom/press-releases/2017/estimates-idaho.html
3. Cao, Jiguo, director. *Fun Data Science*. *YouTube*, YouTube, 8 Jan. 2019, https://www.youtube.com/channel/UC1Wh20PhCEOnrEkk58WweFA/videos. Accessed 6 Aug. 2022.
4. Cao, Jiguo. “Caojiguo/FDACOURSE2019: Slides and R Codes for Functional Data Analysis Course.” *GitHub*, 26 Sept. 2019, https://github.com/caojiguo/FDAcourse2019.
5. “Certain Medical Conditions and Risk for Severe Covid-19 Illness.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 13 May 2021, www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html
6. Data, Health. “Covid-19 Reported Patient Impact and Hospital Capacity by State Timeseries: Healthdata.gov.” *COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries*, 1 Jan. 2020,

https://beta.healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh/data

7. Division, Geography. “Census Regions and Divisions of the United States U.S. Census Bureau.” *Census Regions and Divisions of the United States*, U.S. Department of Commerce Economics and Statistics Administration U.S. Census Bureau, https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

8. Iantsuising. “IANTSUISING/FDA.” *GitHub*, 20 Sept. 2016, https://github.com/iantsuising/FDA

9. National Center for Health Statistics. “Technical Notes: Provisional Death Counts for Coronavirus Disease (Covid-19).” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 29 Apr. 2022, https://www.cdc.gov/nchs/nvss/vsrr/covid19/tech_notes.htm

10. National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases. “About CDC COVID-19 Data.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 17 June 2022, https://www.cdc.gov/coronavirus/2019-ncov/covid-data/about-us-cases-deaths.html

11. Nazario, Brunilda. “Coronavirus Treatment: At Home, Hospital, Drugs.” *WebMD*, WebMD, 26 Jan. 2021, www.webmd.com/lung/covid-treatment-home-hospital#2

12. Nazario, Brunilda. “How Coronavirus Is Transmitted: Here Are All the Ways It Can Spread.” *WebMD*, WebMD, 19 Apr. 2021, www.webmd.com/lung/coronavirus-transmission-overview#1

13. Ramsay, James, et al. *Functional Data Analysis with R and MATLAB*. 2009th ed., Scholars Portal, 2009.

14. Studio, R. “Download the RStudio Ide.” RStudio, 2022, https://www.rstudio.com/products/rstudio/download/

15. “Symptoms of Covid-19.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 22 Feb. 2021, www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

16. “Test to Determine If You HAVE Covid-19 (Swab Test).” *Labcorp*, 2021, www.labcorp.com/coronavirus-disease-covid-19/individuals/infection-test?utm_source=google&utm_medium=cpc&utm_campaign=Labcorp%2BCOVID-19%2BIndividuals%2B-%2BNon-Branded%3BS%3BCE%3BBR%3BINF%3BCO%3BCO%2B%5BE%5D&utm_content=Infection%2BTest&utm_term=covid%2B19%2Binfection%2Btest&gclid=Cj0KCQjwraqHBhDsARIsAKuGZeHisWMOYBV_8bM1wED_TSg-OrHAA66t0ynFKF38qqaOP7lcWJqUiQUaAhUbEALw_wcB&gclsrc=aw.ds

17. “United States COVID-19 Cases and Deaths by State over Time.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 31 July 2021, data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data

18. “United States.” *Worldometer*, 2 Aug. 2021, www.worldometers.info/coronavirus/country/us/