**Predicting and Analyzing the Spread of COVID-19 over Space and Time in the USA**

**Michael Lynch**

**April 12, 2020**

**Introduction**

In this project I have analyzed the spread of COVID-19 over the United States. In addition, I try to predict the increase in number of cases tested positive into the near future (~10 days). This is an important problem because as I write this report, we are in the midst of a global pandemic concerning the novel coronavirus, COVID-19.

**Data**

In the project I have used three datasets. All datasets are currently being updated each day. The first dataset is taken from <https://covidtracking.com>. I use the first dataset to plot the rise in number of cases over time in the state of Minnesota, which is where I live. Then I repeat the process for all 50 states, using a series of for loops. To try and predict the rise in cases into the near future I use a machine learning algorithm called SVM. I won’t get into the specifics of SVM here, but it is a widely used machine learning algorithm. After I use SVM to attempt to predict the rise in cases over the state of MN, I repeat the process for all 50 states, again using a series of for loops.

The second dataset is obtained from the John Hopkins University Center for Systems Science and Engineering GitHub repository at <https://github.com/CSSEGISandData/COVID-19>. This dataset contains stats for each county in each state of the US for the number of cases of coronavirus cases each day. I use this dataset to run k-means clustering to try and find the clusters of areas with cases high in COVID-19.

The third dataset is from the NYTimes’ GitHub repository at <https://github.com/nytimes/covid-19-data>. This dataset contains information similar to the second dataset, except it doesn’t give the geospatial coordinates for each county in the dataset. I use geocoding to get the coordinates and then make a heatmap of COVID-19 cases over the entire US, followed by an animation of COVID-19 spreading over the US from the end of Jan. 2020 until the present (mid-April 2020).

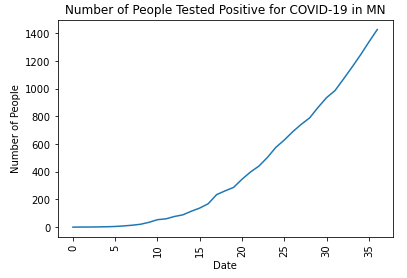
**Methodology**

I use the SVM machine learning algorithm to attempt to predict the increase in number of cases for each state in the entire US. To find the best hyperparameters I run a special function included in the scikit-learn library called “RandomizedSearchCV” to do this. To prove the function did a good job at finding the best hyperparameters I plot the SVM degree number vs. the mean absolute error and we see that the best degree for the data used occurs at degree 3, which is the one chosen by the function. It should be noted that the SVM hyperparameter tuning is run for the state of MN. I use the same hyperparameters for every other state in the US when I plot the predicted number of increase in cases for those states. I realize this is a major assumption and the results would most likely turn out different if the hyperparameters we tuned for a different state, such as New York.

With the second dataset I run a k-means clustering algorithm to try to find the place in the us with the most concentrated cases of COVID-19. The center of the clusters should correspond to places with a large number of cases. The same thing could be done in other ways, such as a heatmap and to show that I later plot a heatmap with a different dataset.

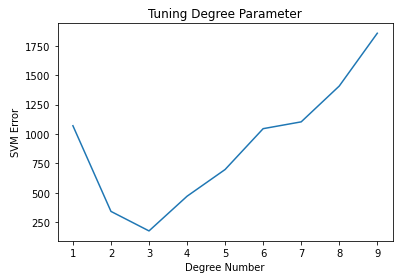
**Results**

In the notebook, the first thing I do after preprocessing the data is plot the number of cases of COVID-19 in my home state of Minnesota. This gives an idea of how quickly the virus is spreading over the state. It can be seen that the slope of the plot continues to increase, which is not a good sign. As stated in the news recently, the objective is to “flatten the curve” and a “flat curve” would be equivalent to the slope in the plot no longer increasing (i.e. an inflection point).



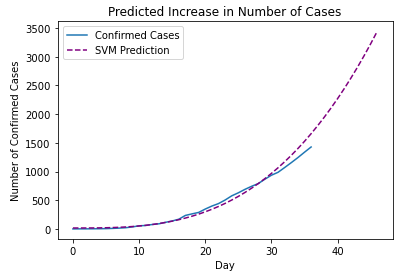
Figure

After getting a basic plot of the number of cases in MN over time, I prepare the data to run through the SVM algorithm and attempt to find the best hyperparameters. After the best hyperparameters are found I plot one of the hyperparameters (degree) versus the error. It can be seen that the degree should be chosen as 3 since that gives the lowest error.



Figure

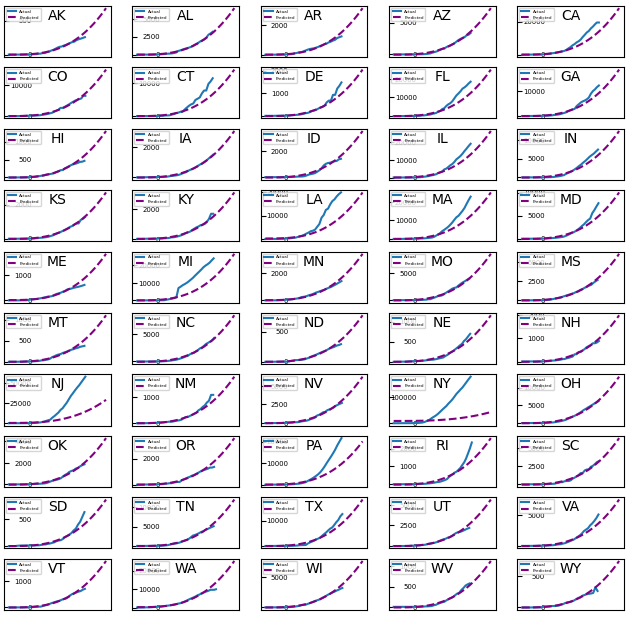
Next, I run the SVM algorithm with the best hyperparameters and then plot the results against the first plot that I made. We can see the the SVM algorithm predicts the number of cases will continue to increase, at least for the next 10 days.



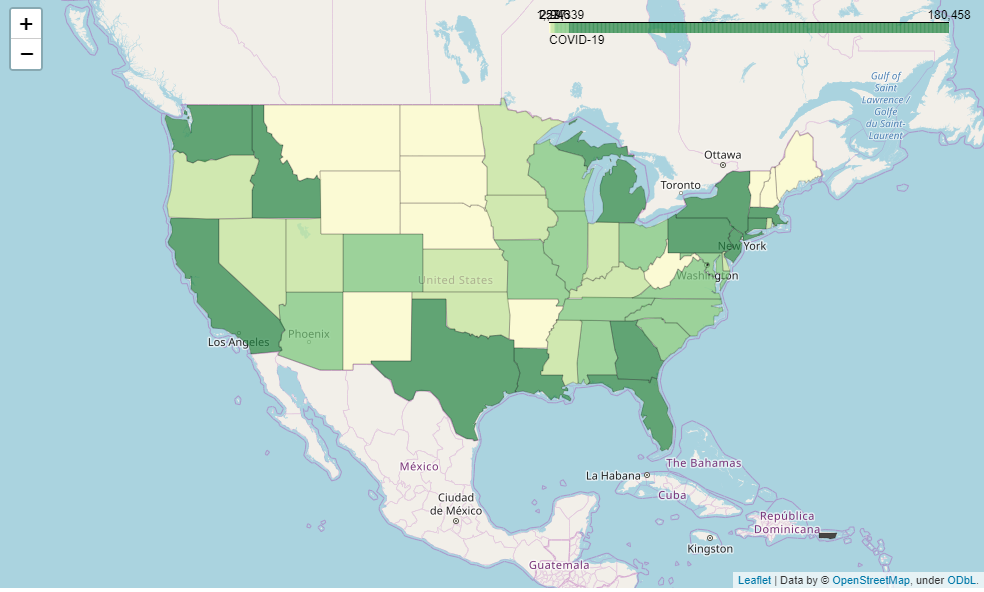
Figure

After predicting the increase in coronavirus cases over time in the state of MN, I repeat the same process except with all 50 states using a for loop. This shows the power of processing data in the Python programming language. If you know how to do something for one object, you can accomplish the same thing for a million object, just using a for loop. Below are the results for each state in the US, followed by the predicted increase in each state. It is interesting to compare the shapes of the curves between each state. It is clear that some states are increasing in numbers faster than others, for example NY compared with ME.

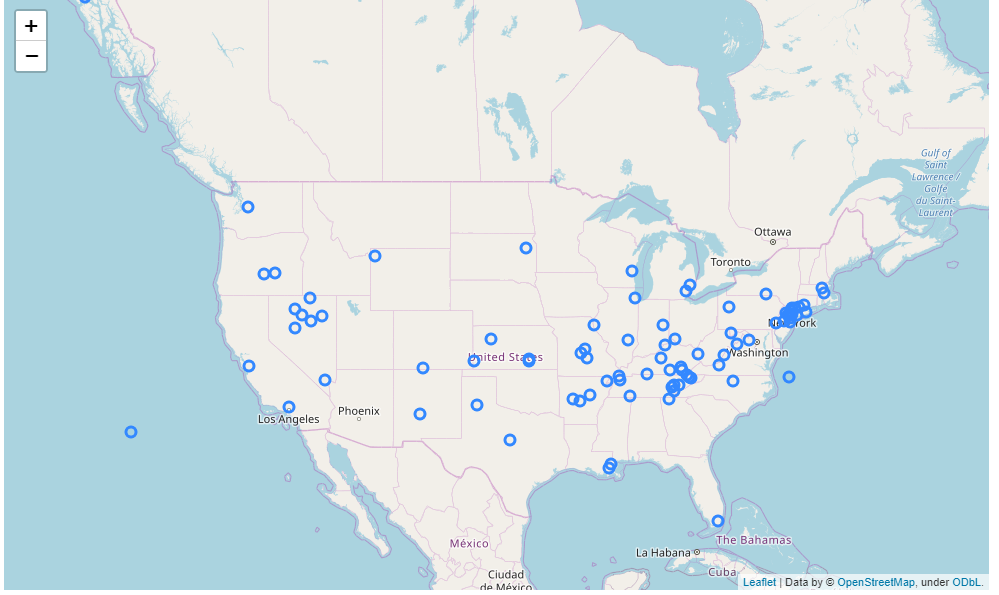
Figure



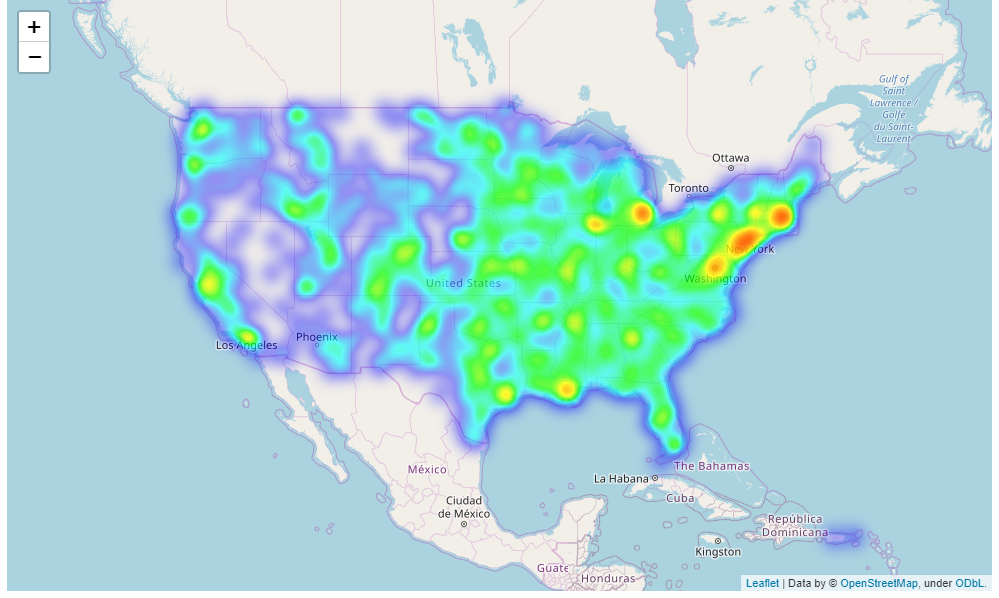
Figure

After predicting the increase in number of coronavirus cases I make a choropleth map using the same data and the folium library of Python. Note that in a lot of cases, states that have the highest number of cases according to the choropleth map, have the steepest curves in the plot of predicted increase of number of cases.

Figure

Using the second dataset I run k-means clustering algorithm and map the centers of the clusters onto a map of the US. Comparing the cluster centers to the choropleth map, we can see that on average the cluster centers lie in states with the highest number of cases, which is what we would expect.

Figure

Lastly, I use geospatial data for each county in the US and make a heatmap of the number of COVID-19 cases in the US. We can compare this heatmap to the choropleth map and the k-mean cluster map.

Figure

**Discussion**

It is clear from the above discussion that the number of coronavirus cases will continue to increase in every state of the US. It is also clear the areas with larger populations have a larger number of cases, which is what would be expected, unless drastic quarantine or stay-at-home orders were issued in those areas.

**Conclusion**

Using Python and various libraries for Python I have analyzed and shown visualization of the spread of coronavirus over the US in both space and time. It is clear from the above analysis that the situation is serious, and the spread isn’t stopping anytime soon.