Covid-19 or corona virus started its spread from China at the end of Year 2020. By February 2020, cases start to appear in the United States and spread rapidly across all the states. The growth in the number of cases has been exponential like any other case of infectious disease. Scientists and researchers around the world are actively looking at the data to understand this virus. In uncertain times like these, it is important to know the factors that affect the spread of this virus to make plausible predictions about the future.

Since the first case of covid-19 was discovered in the US in Snohomish county in the Washington State, the virus has now spread to other places and in some cities like New York and Chicago, the spread has been faster, making them new hotspots for the virus even though cities like Seattle & Los Angeles reported their first cases before these cities. A simple line graphs for some counties that saw their first case of infection around the same time shows the difference in the rate of virus spread.

This poses a question that; what are the different factors that are leading to differences in the pace of covid-19 spread across US counties? Can a cross sectional comparison of these counties help us identify any socio-economic or geographical feature that has any effect on virus spread and can we use these features to predict the number of infections for other counties that are behind in trend. We carry out a cross sectional analysis of counties and their features in comparison to the number of covid-19 cases reported by using statistical and data mining tools. The purpose of the report is to identify any trend or characteristics of counties that are connected to the pace of virus spread.

Data

* Cases & Deaths: We collect the data of total number of covid-19 infection cases and deaths at county level as on April 30, 2010. The data has been obtained from NYTimes Github repository.

We selected 1462 counties which had non-zero number of deaths reported. This was done because ACS data was only available for counties that have any death reported for covid-19.

We divided the total number of cases by population to have cases per hundred thousand of population. This gives a good comparison of cases across counties with different population size. From now on we will refer to cases as cases per hundred thousand of population. The density graph shows that a lot of counties have lower number of cases and a very skewed distribution of cases.

* Days since first Case: We calculated the total number of days since 1st infection for all counties as on April 30, 2020. This variable measures the time period for spread of virus.
* Days since Stay at home orders: Similarly, we calculated number of days since stay at home orders were issued by State and County authorities as on April 30, 2020. This gives us the time period for interventions taken by authorities. Along with this we also obtained Google’s movement analytics for each county which gives a percentage decline in people’s movement from the baseline average.
* Socio-Economic Variables: We obtained socio economic variables on the population of counties from American Community Survey reports. Main variables include: Unemployment rate, % of White, % of Households in Poverty, % of population employed in different industries, median income and % of population with a bachelor’s degree.
* Demographic Variables: % of male, % of age between 18 and 64, % of White, Median age, % of total households as family etc. The complete list is attached in the appendix.
* Geographical variables: We obtained variables like population density, average of last 8 years highest summer temperature and humidity rate. We wanted to obtain the recent monthly averages for counties but could not find the latest data, so we used these variables. A look at US map shows that cases are higher in highly dense counties, especially on the coast. This can also be because these counties are closer to big cities that see more international traffic and might be the first one to get infected by virus.

There were many socio-economic variables that were very highly correlated, so we dropped some variables and finalized 30 variables for this analysis. The complete list of these variables is provided in the appendix.

Before we start with our analysis, it is important that we acknowledge that number of cases reported depends heavily on the number of tests conducted. Unfortunately, testing policy has not been uniformed across US and that can lead to a measurement error in cases. Secondly, there are reports that the given time frame for start of covid-19 is not accurate and it is presumed that covid-19 reached many major cities of the US way before the first case was reported.

However, due to data limitation, we must assume that the official reporting of cases represents somewhat actual position for the start of virus infection. For testing, we tried to get data on total tests conducted on county level but only find the data at state level. So, we calculated total tests conducted per hundred thousand people for each state and then use it for each county in respective state. This is not perfect, but we are assuming that the rate of testing is uniform across counties in each state and will help in differentiating counties of different states based on testing rates.

Principal Component Analysis

We use principal component analysis to see if we can look at the variations in the counties and evaluate if those variations are related to number of cases in these counties. After scaling the data, we see that 90% of variation in 30 explanatory variables can be explained by 16 principal components, pointing towards correlation between these variables.

The biplot of main variables that contribute to variance is first two PCs shows that PC1 captures variation in population density, time since first case, proportion of population between 18 and 64 and change in the movement of people to their workplaces.

Plotting the scatter plot with first two components with color scaling the number of cases, we see that there is no specific clustering of high cases counties but generally, we see that there are more red points in the upper left quadrant of the graph. Which means that variables mentioned above are correlated with higher number of covid-19 cases in counties. Similarly, lower right quadrant has less number of high case counties and from the biplot we saw that this is associated with higher proportion of white population, higher median age and higher proportion of spouse households in total households. Interestingly, we cannot say much about the effect of temperature on virus spread but there does seems to be a slight negative correlation between temperature and virus spread.

Negative Binomial Logistic Model

Since our outcome is count of cases with overdispersion, we fit a negative binomial logistic model on our data. As we have only those counties that reported the any cases, we use zero truncated negative binomial logistic model. This model has been used by \_\_\_\_\_\_\_\_\_ in their study of air quality exposure’s effect on covid death rate in US counties. We include all the explanatory variables in our model which helps us ascertain the correlation of these variables with cases. The results are provided as under:

We can clearly see that infection start date is a very important variable as expected. Along with it, stay at home orders are linked negatively to cases growth which is also theoretically correct. Another important variable is population density per square mile area which means that congested counties see higher number of cases.

Counties with higher percentage of educated people and people with race as white also observe lower cases. Higher proportion of people employed in educational and health services and lower proportion of people in arts, recreational, accommodation and public administration industries are correlated with higher cases of virus. The signs for these co-efficient are consistent with our PCA analysis.

We have a problem of small sample and it would have been better if the sample size could have been bigger. We have Hauck-Donner effect in two variables: % of population age 18 to 64 and E\_Total\_Households\_TYPE\_Family. The co-efficient magnitude cannot be considered as casual effect. However, the signs of co-efficient points towards some plausible correlation between number of cases and different variables.

To check the predictive power of our model, we do 300 test/ train splits of our sample and run the model. We get a RMSE of around 240 cases per hundred thousand people. Plotting the predicted values against the actual values, we see that the model on average underpredicts the number of cases for counties with high number of cases and overpredicts for counties with lower number of cases. Overall, we do not think this model does a good job at prediction.

We try the same model with actual number of cases rather than number of cases per hundred thousand people and include log of population in the model to control for community size. The RMSE of 300 test and train split comes out to be around 1500 and the prediction graph is more balanced as compared to the model with cases per hundred thousand population. Therefore, we prefer using total number of cases in the model for prediction purpose.