Covid-19 Spread - Cross sectional Analysis of US Counties

Muhammad Zargham, Clark Granger, Zachery Carlson

Abstract

We conduct a cross sectional analysis of Covid-19 cases as on April 30, 2019 across the US counties based on their socio-economic and geographical factors that may be important in explaining the different pace of virus spread. Along with identifying these important factors, we try to fit negative binomial, KNN and Random forest models to predict the number of covid-19 cases and to test whether incorporating these socio economic variables into our models can lead to better predictions of covid cases at county level. We find that some important socio-economic features are correlated with the number of cases and can give further insight for studying this pandemic.

Contents

Introduction

Covid-19 or corona virus outbreak started from China at the end of Year 2020. By February 2020, cases start to appear in the United States and spread rapidly across the country. The growth in the number of cases has been exponential like in the case of any highly contagious disease. In uncertain times like these, as scientists and researchers around the world are actively working to come up with a cure, it is important to know the factors that affect the spread of this virus to slow down its spread and make plausible predictions about the future.

Since the first case of covid-19 was discovered in the US in Snohomish county of 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 quite earlier than others, we see a slower pace of outbreak in 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.

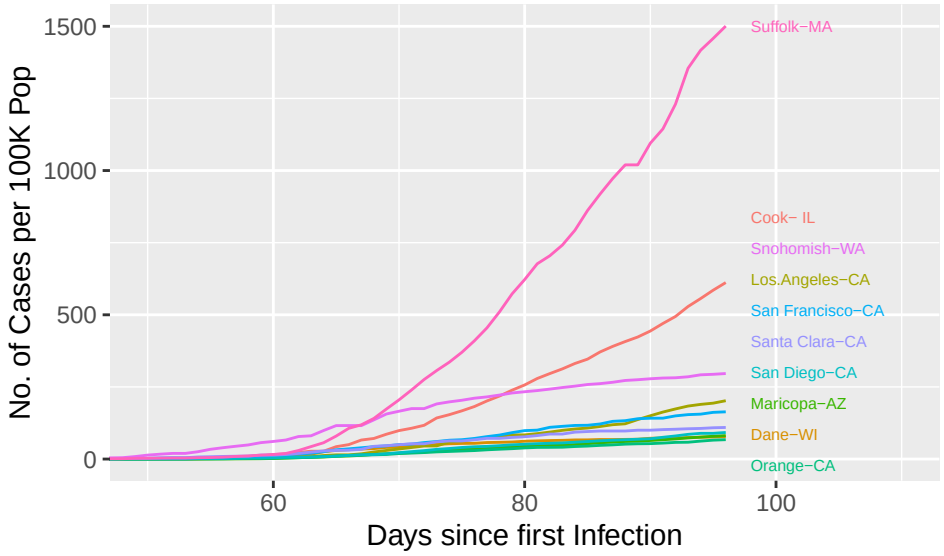


Figure 1: Trend Line Graph for Counties

This poses a question: what are the different factors that are leading to differences in the pace of covid-19 spread across the US counties? Can a cross sectional comparison of these counties help us identify any socio-economic or geographical features that have effect on virus spread and can we use these features to predict the number of infections for other counties that are behind the 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 NY-Times Github repository. We selected 1462 counties which had non-zero number of deaths reported. This was done because the 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 populations. 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.

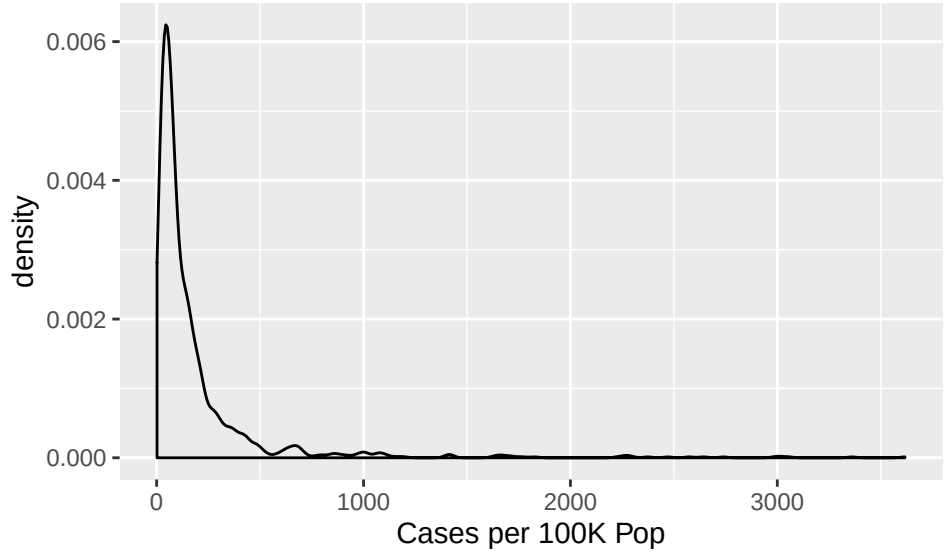


Figure 2: Density Graph for Number of Cases

- **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. A scatter plot of days since first infection and number of cases per 100 thousand people shows that there is a general exponential trend but some counties have been able to keep their number of cases down even though the got the virus before other counties.

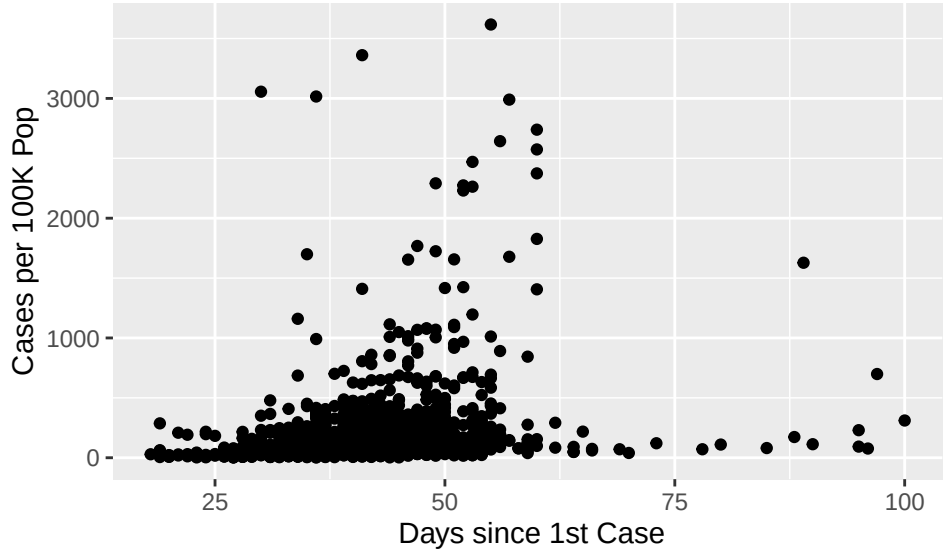


Figure 3: Scatter Plot of Time and Cases

- **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. This information was obtained from Jie Ying Wu github respository. 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 available on github repositories for Covid data. 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 also gathered 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.

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.

Note: 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 found the data at the state level. So, we calculated total tests conducted per hundred thousand people for each state and then use it for each county in the respective state. This is not perfect because we are assuming that the rate of testing is uniform across counties in each state but we think that public health and testing policy is same across the all counties in each state.

Methods

Principal Component Analysis

We use principal component analysis to summarize the variations in correlated socio-economic factors and see if it explains differences in covid spread. 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 for 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 tight clustering of high case 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. 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 humidity level 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 any covid cases, we use zero truncated negative binomial logistic model. This model has been used by Wu,Nethery,Sabath (2020) in their study of exposure to air pollution and its effect on covid death rate in US counties. We include all the explanatory variables in our model to see their marginal effect on log case count. The summary of the model is provided in the appendix.

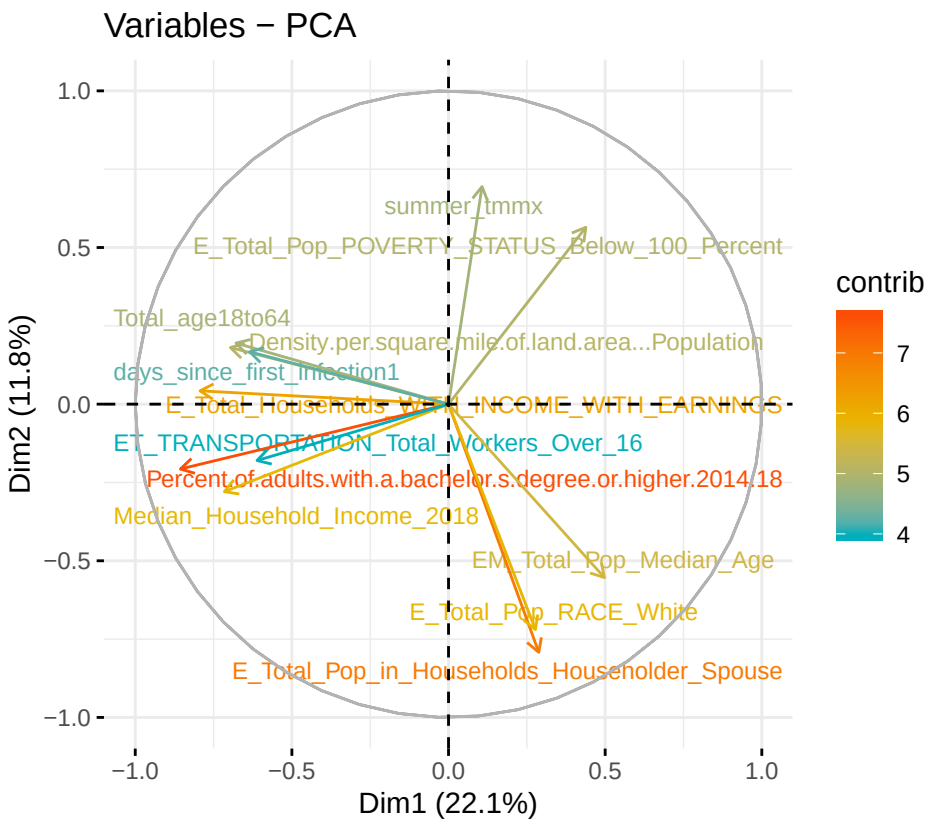


Figure 4: Variables Contribution in first two PCAs

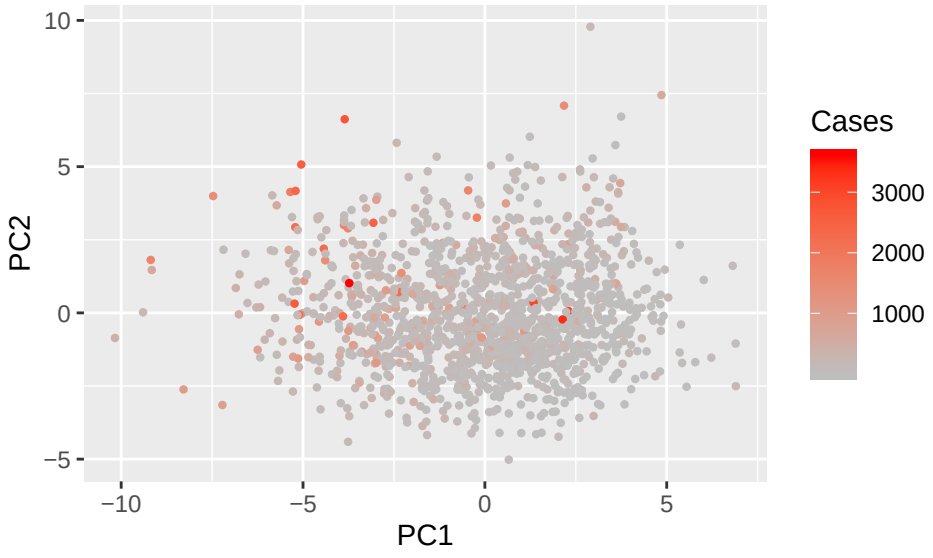


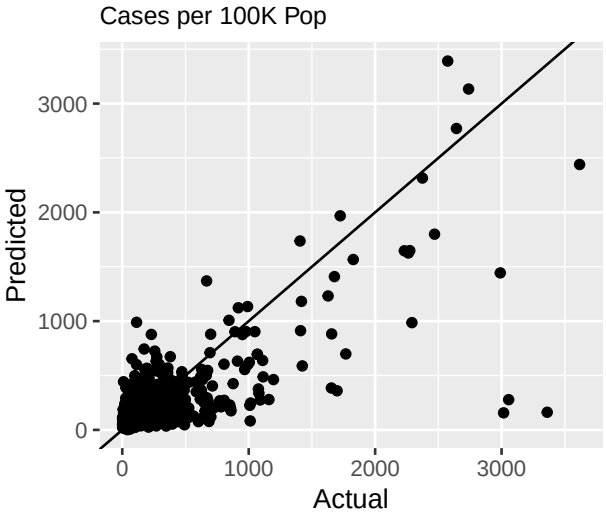
Figure 5: Scatter Plot of Cases along Principal Components

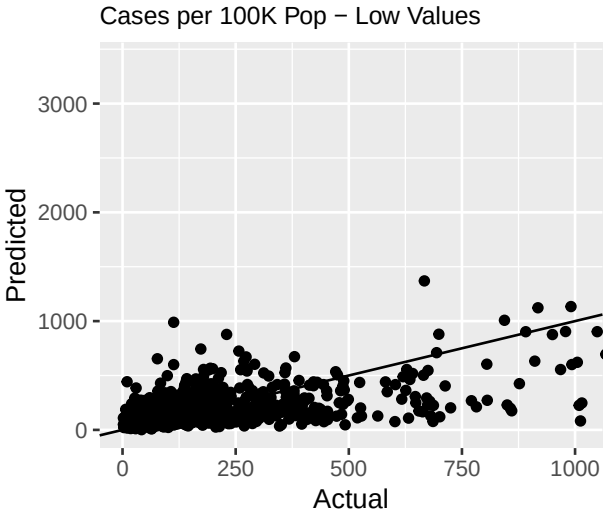
We can clearly see that infection start date is positively correlated with the log count of cases. Along with it, days since stay at home orders are linked negatively to cases growth which is also theoretically correct. Another important variable is population density per square mile area which means that congested counties see higher number of cases.

Counties with higher percentage of educated people and white people also observe lower cases. People fitting the above profile are more likely to be rich, living in large and open areas which in turn increases social distance and lowers virus transmission. 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.

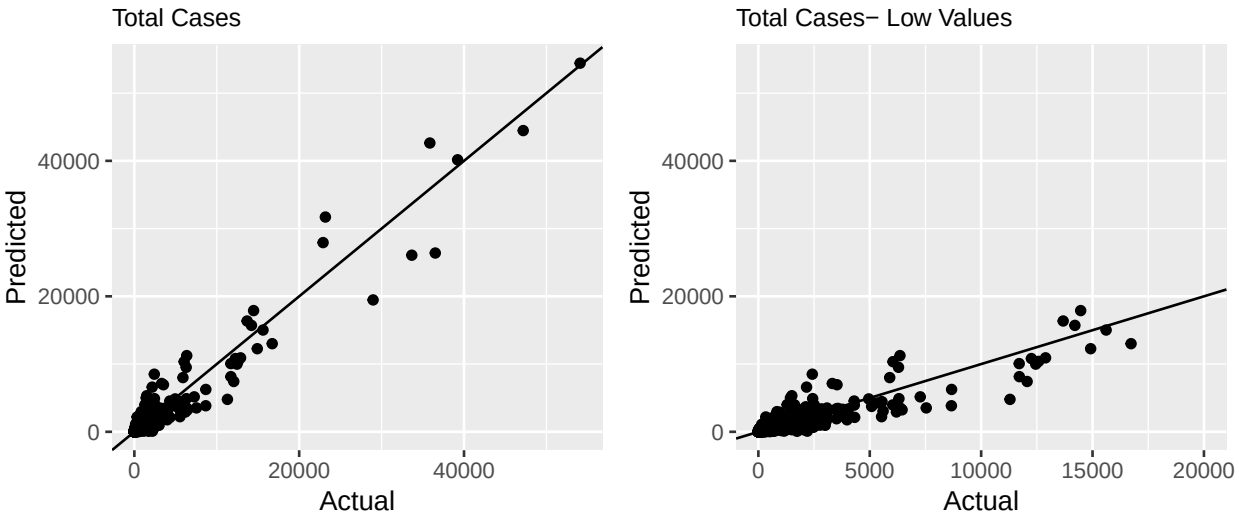
We have a problem of small sample and it would have been better if the sample size could have been bigger. We have Hauck-Donner effect in two variables: % of population age 18 to 64 and E_Total_Households_TYPE_Family. The co-efficient magnitude cannot be considered as casual effect. However, the signs 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 would prefer using total number of cases in the model for prediction purposes.



K-Nearest Neighbors

We try k-nearest neighbors technique to predict the number of cases for counties.We use total number of cases instead of cases per 100K population. We first use PCA to scale the data and reduce the number of variables. Using 9 principle components, we select K = 3 from the K-RMSE plot. The knn model gave us an average out of sample RMSE of 1499. The RMSE for knn model is around the same level we obtained from negative binomial model. However, this model does not provide any insight into the importance of variables in determining the number of cases.

Analysis using Random Forest and Boosting Trees

We fit random forest and boosting models on cases per 100k population. By using these models we can also get information about variable relevance .After a train test split of our sample, we created 1000 trees by random forest including a minimum of 5 features per bucket.

Table XX (APP TABLE) summarizes the variable importance information derived from the random forest model (RF). We can notice that as expected the population density per square mile plays an importante role as expected. It suggests that Counties more dense must expect a higher number of people infected, as happened in New York state. From the variable importance table we can also see demograohic and economic covariates of population that are more likely to be exposed to the virus, such as occupation type. As mentioned, we fit a Boosting Tree(BT) that allows us to validate the results of the RF model. The BT model generates a variable importance output as well, the result in this case are closed with those released by the RF and are summarized in table XX(APP TABLE).

From Random Forest model and Boosting Trees we have the posibility to generate partial dependence functions. This functions show the marginal effect of one feature on the predicted outcome of our model. A partial dependence function can be represented as:

$$\hat{g}_{x_i}(x_i) = E_X\left[\hat{g}(x_i, X)\right] = \int \hat{g}(x_i, X) d\mathbb{P}(X)$$

Where x_i are the variables for which we are computing the function and matrix X the other variables used in the model $\hat{g}$. Then, the partial function tell us for a given value of covariate i what the average marginal effect on the prediction is. Then by the calculation of the partial depence functions we can look for the effect of the County level features over the number of Covid cases per 100k. Figure **??** shows the partial depence function derived of the RF for some features we found intuitive, for example we can observe the can confirm the positive relation of population density over number of cases. We also can evaluate the effect of a

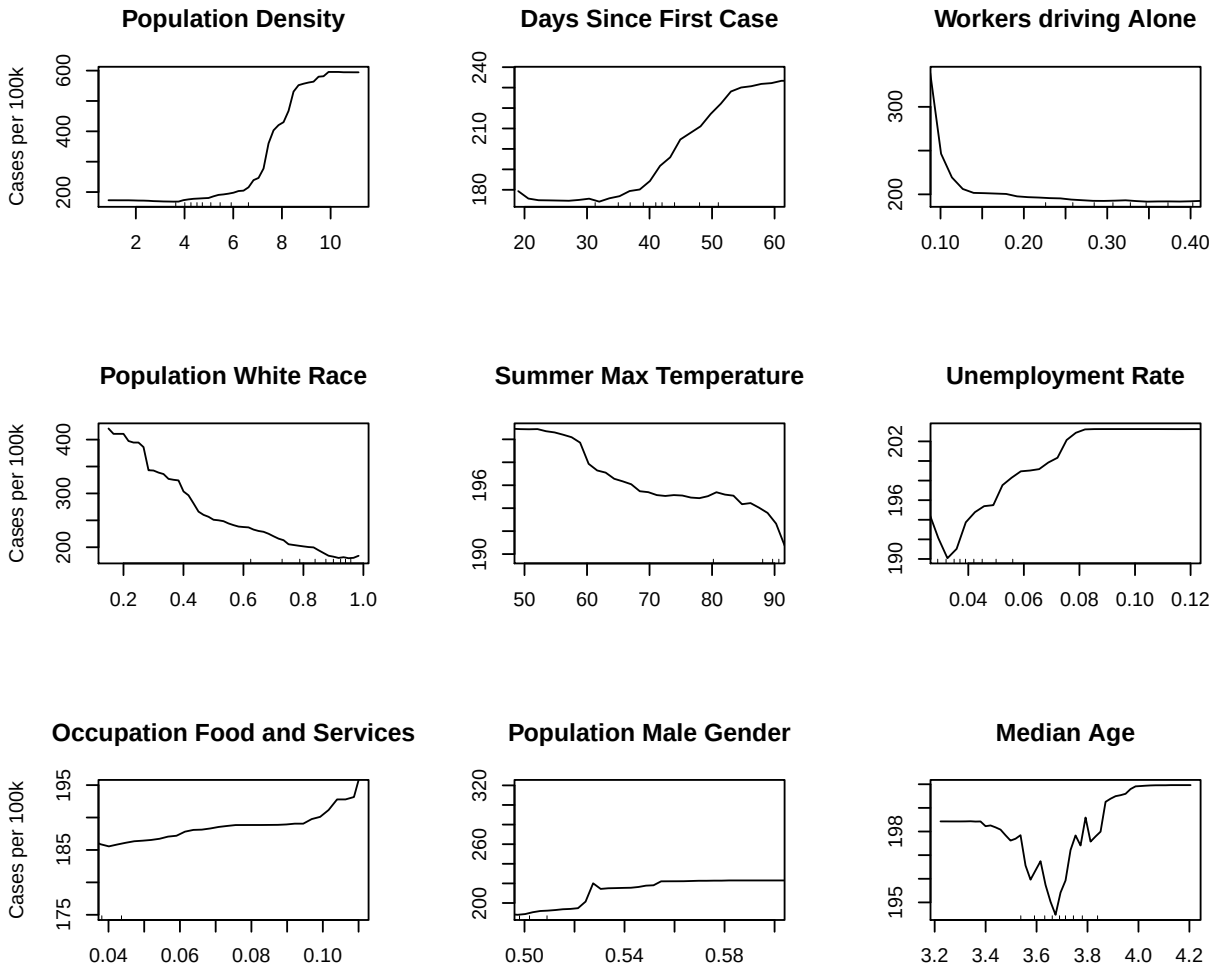


Figure 6: Partial Dependence From Random Forest Model

socioeconomic variable such as poverty level, specifically we can see that an increase of poverty level has a positive impact over number of cases. In the case of proportion of population being white has a negative effect on number of cases. This finding goes in line with the last reports of New York State that suggest that the most affected ethnicity or race are Hispanic and Black. It is also interesting to see how warmer summer temperatures have a negative effect over Covid cases. Finally, we can see how a higher median age has a positive effect and more people driving alone to work has a negative one.

Regarding the predictive performance of tree models, we got an out of sample RMSE of 247 for the RF model and 301 for the BT. In case of total cases, we get an out of sample RMSE of around 1400 which is slightly less than negative binomial and KNN models.

Conclusion

By running different models on the available data, we were able to highlight some county level factors that are correlated with the spread of covid-19 virus. From our analysis, we have gathered that high population density, lower level of education, higher unemployment, higher median age, lower proportion of white race and male and more employment in food, accomodation and health sectors are factors that are correlated with higher number of cases of covid-19 per 100 thousand people. All of these models show similar predictive power but we think it can be further improved by including more counties and including testing data for county level. With the given data, we conclude that random forests technique gives us the best predictions.

Appendix

List of Variables

	Name	Description
1	fips	Unique County Code
2	county	County Name
3	State	State Name
4	cases	Total No of Cases as on April 30, 2020
5	deaths	Total No of Deaths as on April 30, 2020
6	days_first_case	Total Days since reporting of 1st case
7	Tests_per_100K	Total Tests conducted per 100K of Population at State Level
8	Stay_home_order	Total Days since issuance of stay at home orders
9	Adults_Bachelor_degree_or_higher	% of Adults with bachelors or higher degree
10	Unemployment_rate_2018	Unemployment rate
11	Population_age_18to64	% of Population between age 18 & 64
12	Total_Pop_SEX_Male	% of Population male
13	Total_Pop_RACE_One_Race	% of Population belonging to one race only
14	Total_Pop_RACE_White	% of Population belonging to white race
15	Total_Pop_in_Households_Householder_Spouse	% of Population living in households with spouse
16	Total_Pop_in_Households_Householder_Parent	% of Population living in households with parents
17	Total_Pop_Over_15_Divorced	% of Population over 15 divorced
18	Occ_Administrative_Waste_Management	% of employment in administrative & waste management
19	Occ_Educational	% of employment in educational services
20	Occ_Healthcare	% of employment in heath care sector
21	Occ_Arts_Entertainment_Recreation	% of employment in arts and recreational sector
22	Occ_Accommodation_Food_Services	% of employment in accomodation and food services
23	Occ_Other_Services	% of employment in other services
24	Occ_Public_Administration	% of employment in Public administration
25	Total_Workers_Over_16	% of Total Population above 16 and working
26	Total_Workers_Over_16_Drove_Alone	% of Total Population above 16 and working that drove alone to work
27	Total_Households_TYPE_Family	% of total households categorized as Family
28	Households_with_incomes	% of total households with incomes
29	Total_poverty	% of Population in Poverty
30	summer_tmmx	8 Year average maximum summer humidity
31	summer_rmax	8 Year average maximum summer temperature
32	Population_density_per_sqrm	Population density per square mile of area
33	Median_Age	Median age (log)
34	Median_Household_Income	Median household incone (log)
35	Total_Population	Total Population of County (log)
36	workplaces_percent_change_from_baseline	Total % decrease in movement to workplaces from the baseline. Based on daily average of 20-29 April
37	cases100k	Log count of total Cases

Summary of Negative Binomial Model

vglm(formula = cases100k ., family = pospoisson(), data = covid3)				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-34.95	0.55	-64.10	0.00
days_first_case	0.01	0.00	26.92	0.00
Tests_per_100K	0.00	0.00	148.86	0.00
Stay_home_order	-0.01	0.00	-47.32	0.00
Adults_Bachelor_degree_or_higher	-4.19	0.05	-79.67	0.00
Unemployment_rate_2018	-2.80	0.25	-11.00	0.00
Population_age_18to64	-4.56	0.13	-35.58	0.00
Total_Pop_SEX_Male	12.08	0.21	58.44	0.00
Total_Pop_RACE_One_Race	13.42	0.21	64.49	0.00
Total_Pop_RACE_White	-1.94	0.02	-102.95	0.00
Total_Pop_in_Households_Householder_Spouse	-1.64	0.05	-30.81	0.00
Total_Pop_in_Households_Householder_Parent	9.42	0.59	16.06	0.00
Total_Pop_Over_15_Divorced	-7.38	0.21	-35.98	0.00
Occ_Administrative_Waste_Management	-10.70	0.59	-18.13	0.00
Occ_Educational	2.21	0.22	9.87	0.00
Occ_Healthcare	7.48	0.18	40.92	0.00
Occ_Arts_Entertainment_Recreation	-5.15	0.50	-10.23	0.00
Occ_Accommodation_Food_Services	-18.92	0.34	-56.05	0.00
Occ_Other_Services	-25.74	0.68	-37.89	0.00
Occ_Public_Administration	-2.61	0.20	-12.86	0.00
Total_Workers_Over_16	-1.23	0.06	-21.31	0.00
Total_Workers_Over_16_Drove_Alone	1.97	0.06	31.39	0.00
Total_Households_TYPE_Family	-3.07	0.07	-41.66	0.00
Households_with_incomes	7.01	0.09	79.18	0.00
Total_poverty	6.39	0.10	63.86	0.00
summer_tmmx	-0.01	0.00	-11.23	0.00
summer_rmax	-0.00	0.00	-5.59	0.00
Population_density_per_sqrm	0.21	0.00	76.21	0.00
Median_Age	2.73	0.05	55.78	0.00
Median_Household_Income	1.35	0.03	51.87	0.00
workplaces_percent_change_from_baseline	-5.41	0.05	-118.65	0.00
Name of linear predictor	loglink(lambda)			
Log-likelihood	-96986.51 on 1411 degrees of freedom			
Number of Fisher scoring iterations	10			
Hauck-Donner effect	(Intercept) Population-age-18to64 summer-tmmx			