Ahmed Boutar

Scott McKinney

Understanding the Environmental Factors that Contribute to Spread of COVID-19

University of Rochester

Abstract

We are in the middle of a global pandemic that has basically brought the world to a halt and claimed the lives of thousands of people. Coronavirus is incredibly contagious and has spread at an unprecedented rate. The goal of our project was to better understand the local factors which most contribute to the spread of the virus in an area.

We manually collected data from multiple sources to build a dataset which contains information on many factors for different locations. We determined which factors these will be by looking at things that usually contribute to the spread of other infectious diseases. We then trained a multivariable linear regression model on this data to find the importance of each attribute. We used this to determine which of the factors contribute most to the spread of COVID-19. For variables which were assigned high importance or piqued our interest, we utilized basic data visualization techniques to further explore the data. Hopefully, in doing this, we can gain insight into what features of a region support and hurt the spread of COVID-19.

1. Introduction

1.1 Background and Motivation

The Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus, a family of viruses named after their spiky crowns. This virus caused a global pandemic and has basically brought the world to a halt and claimed the lives of thousands of people. This virus is incredibly contagious and has spread at an unprecedented rate. The first human case of the virus (named SARS-CoV-2) was first reported by officials in Wuhan City, China, in December 2019 [1]. Since then, over 3.7 million cases have been registered as well as over 260,000 deaths world-wide.

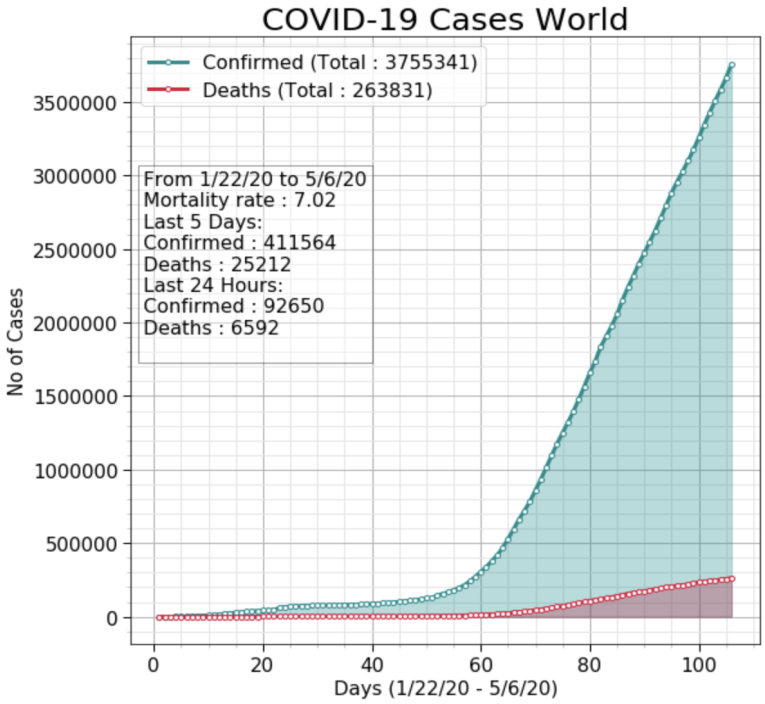


Figure 1. COVID-19 Cases World. Reprinted from COVID-19 Case Study - Analysis, Viz & Comparisons, In Kaggle, by T.K, Retrieved May 08, 2020, from https://www.kaggle.com/tarunkr/covid-19-case-study-analysis-viz-comparisons

The COVID-19 virus spreads primarily "through droplets of saliva or discharge from the nose when an infected person coughs or sneezes" [1]. It can cause mild to severe illness, with severe illnesses occurring in adults 65 years and older as well as people with underlying medical problems [2].

As this is a global pandemic, there are a lot of data science efforts out there focusing on modeling the spread of this disease. Most of these, however, focus primarily on modeling the spread of the disease with the goal of estimating the number of patients that will be infected as opposed to determining which factors are most important in causing that spread. Those which do focus on the factors causing the spread look at individual factors and how they affect the spread as opposed to looking at the combination of factors and comparing those factors to each other. Therefore, we decided to work on determining the factors that most contribute to the spread of the virus in a chosen geographic location.

1.2 Problem Statement

"What are the factors that most contribute to the spread of COVID-19?"

We will collect data from multiple sources to build a dataset which contains information on multiple factors for different locations. We will determine which factors these will be by looking at things that usually contribute to the spread of other infectious diseases. We will then train a model on this data to determine which of the factors, or combination of factors contribute most to the spread of COVID-19.

1.3 Related Work

Like mentioned before, most of the work that has been done around the novel SARS-CoV-2 virus mainly focuses on predictions such as predicting the evolution of the spread curve or the different possible ways to flatten the spread curve. Examples of such studies are available on websites such as Kaggle.com and are often the product of the collaboration of multiple parties. Other research focuses on the preventative measures that governments are implementing as well as its implications. For instance, Dantas et. al. (2020) focused on studying the implications of social distancing in Brazil during the pandemic [3]. The authors argue that although social distancing is "[one of] the best measures to protect against the virus", developing countries, due

to their fragile economies, might not be able to sustain such initiatives, and therefore, resulting in further human and economic catastrophes. The authors acknowledge that "the flattening of the COVID-19 curve will require additional measures in developing countries, where the spreading factor of the virus is differentiated and more complex." [3]

Other research focuses more on determining the factors of the spread of the COVID-19. For example, Sean Wei Xiang Ong et. al. (2020) worked on determining the association of a higher BMI with severe coronavirus disease in young patients. By conducting a study of 182 patients, the authors determined that obesity can be "a significant risk factor for the development of severe COVID-19, especially in younger patients aged <60 years old" [4]. However, research that studies the factors that contribute to the spread of the virus is still very limited.

2. Methodology

2.1 Overview of dataset

The database we used in our work was built from scratch. It contains 20 attributes that represent different measures that can be interpreted as factors.

daysSinceInf	county	state	state code	cases	deaths	avg rate of ir	avg rate of d	population	avgTemp	avgPrecip	percentage c	CO 2nd Max	CO 2nd Max	NO2 98th Pe	NO2 Mean 1	Ozone 2nd M	Ozone 4th M	PM2.5 98th I	PM2.5 Weigl	PM10 2nd M	PM10 Mea
29	Mobile	Alabama	AL	594	23	20.4827586	0.79310345	413210	62.5	58.3	19%	.	.	.	.	0.07	0.059	17	8.3	.	.
35	Tuscaloosa	Alabama	AL	132	0	3.77142857	0	209355	62.5	58.3	18%	.	.	.	.	0.07	0.06	15	7.9	.	.
30	Coconino	Arizona	AZ	299	26	9.96666667	0.86666667	143476	58.2	13.6	16%		2.3	.	.	0.07	0.063	.	.	.	.
82	Maricopa	Arizona	AZ	2264	64	27.6097561	0.78048781	4485414	58.2	13.6	14%	2.9	2.3	52	25	0.1	0.074	27	8.4	145	
31	Navajo	Arizona	AZ	410	11	13.2258065	0.35483871	110924	58.2	13.6	21%	2.05	2.3	.	.	0.08	0.066	.	.	.	.
39	Pima	Arizona	AZ	760	37	19.4871795	0.94871795	1047279	58.2	13.6	15%	1.2	0.6	30	7	0.08	0.065	9	5	139	
42	Pinal	Arizona	AZ	197	5	4.69047619	0.11904762	462789	58.2	13.6	16%	2.05	0.6	.	.	0.09	0.072	33	11.3	222	
29	Yavapai	Arizona	AZ	68	1	2.34482759	0.03448276	235099	58.2	13.6	17%	2.05	0.6	.	.	0.07	0.061	.	.	.	.
33	Garland	Arkansas	AR	105	0	3.18181818	0	99386	60.3	50.6	19%		.	.	.	.	.	16	8.8	.	.
47	Alameda	California	CA	1069	40	22.7446809	0.85106383	1671329	57.1	22.2	10%	5.6	1.7	48	15	0.1	0.072	19	9.4	.	.
45	Contra Costa	California	CA	631	16	14.0222222	0.35555556	1153526	57.1	22.2	10%	6.6	1.8	34	7	0.09	0.069	17	7.8	53	
41	Fresno	California	CA	295	7	7.19512195	0.17073171	999101	57.1	22.2	13%	1.8	1.4	63	20	0.1	0.08	37	11.2	234	
28	Imperial	California	CA	143	2	5.10714286	0.07142857	181215	57.1	22.2	13%	4.3	3.1	49	9	0.11	0.08	27	10.8	162	
31	Kern	California	CA	553	3	17.8387097	0.09677419	900202	57.1	22.2	15%	1.2	0.8	56	12	0.1	0.084	47	13	382	
82	Los Angeles	California	CA	10854	455	132.365854	5.54878049	10039107	57.1	22.2	11%	3.8	2.9	78	23	0.12	0.101	30	13.4	159	
51	Marin	California	CA	187	10	3.66666667	0.19607843	258826	57.1	22.2	10%	1.3	0.8	40	8	0.08	0.059	14	6.4	30	
31	Monterey	California	CA	119	3	3.83870968	0.09677419	434061	57.1	22.2	12%	4.5	0.8	26	4	0.07	0.058	12	5.6	76	
46	Placer	California	CA	130	8	2.82608696	0.17391304	398329	57.1	22.2	10%	2.64	1.375	43	7	0.09	0.079	21	7.2	54	
41	Riverside	California	CA	2264	59	55.2195122	1.43902439	2470546	57.1	22.2	12%	1.8	1.2	53	14	0.13	0.096	36	12.5	139	
56	Sacramento	California	CA	879	32	15.6964286	0.57142857	1552058	57.1	22.2	12%	1.7	1.3	55	12	0.1	0.071	30	8.4	90	
33	San Bernardi	California	CA	1032	47	31.2727273	1.42424242	2180085	57.1	22.2	13%	2.2	1.1	74	29	0.14	0.106	34	15.4	126	
67	San Diego	California	CA	2087	63	31.1492537	0.94029851	3338330	57.1	22.2	11%	3.5	1.7	47	14	0.1	0.076	17	8.6	153	
75	San Francisc	California	CA	1022	17	13.6266667	0.22666667	881549	57.1	22.2	10%	1.2	0.9	46	10	0.09	0.053	18	7.7	34	
38	San Joaquin	California	CA	369	17	9.71052632	0.44736842	762148	57.1	22.2	12%	2.2	1.3	58	12	0.09	0.069	34	9.4	116	
34	San Luis Obi	California	CA	125	1	3.67647059	0.02941177	283111	57.1	22.2	11%	2.64	1.375	27	4	0.07	0.07	20	7	129	
46	San Mateo	California	CA	767	28	16.673913	0.60869565	766573	57.1	22.2	9%	2	1	44	9	0.08	0.054	17	7	.	.
33	Santa Barbai	California	CA	354	3	10.7272727	0.09090909	446499	57.1	22.2	11%	2.3	0.6	26	5	0.08	0.068	16	6.8	87	
77	Santa Clara	California	CA	1833	70	23.8051948	0.90909091	1927852	57.1	22.2	8%	2	1.5	52	14	0.09	0.064	25	9.1	53	

Figure 2. Screenshot of the database used in this project

The dataset contains information about different factors that were collected online and were deemed to contribute to the spread of the virus. Since there is a lot of uncertainty about the reporting of cases world-wide, we chose to work with the statistics reported by the United States,

assuming they are the most accurate. We decided to work on the county level since there is a lot of useful data available online and since it gives a more detailed view of the pandemic situation in the United States.

The attributes we used in our dataset are as follows:

- Days since first infection (.i.e. first infected patient)
- The number of registered cases
- The average rate of infection which is calculated based on the number of infections divided by the number of days since the first case. The average rate was calculated separately for each individual county
- The average rate of deaths, calculated based on the number of deaths in a specific county divided by the number of days since the first registered death due to the virus.
- Population of the county, a factor that should be studied knowing that the virus mainly spreads throughout human contact.
- Average temperature, calculated based on the registered temperatures since the first infection was registered
- Average precipitation, calculated based on the registered precipitation data since the first infection was registered
- Percentage of smokers, retrieved online using a web scraper.
- Weather quality: it is represented by multiple quantitative attributes that are proven to represent the quality of the air in a specific county. For instance, the Particulate Matter (PM) is used to describe the mixture of solid particles and liquid droplets in the air.

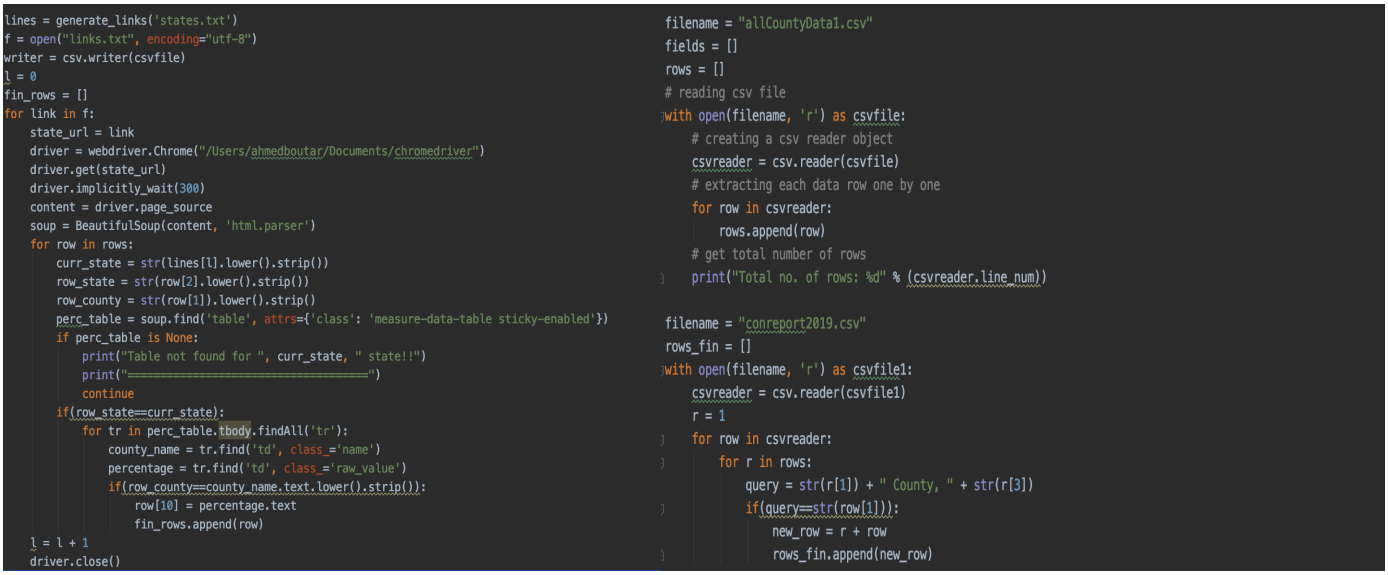
2.2 Data Acquisition

The dataset we created represents different quantitative values of attributes, which represent factors that contribute to the spread of the virus. In the beginning, we had to determine what factors we should focus on in order to understand their influence on the spread rate. We looked at different scientific articles as well as news articles. We gathered factors that were mentioned and found scientific data to back it off. For instance, there is a large amount of research proving that smoking inflames the lungs and suppresses immune function. It also weakens the heart, another risk factor for severe disease [5]. Reducing the functions of the immune system leads to a higher

probability of getting infected (i.e. testing positive for the virus), which makes smoking an indirect factor that contributes to the spreading of the disease, making it a factor of interest in our research.

In addition to understanding how large of a role indirect transmission plays in the COVID-19 pandemic, the researchers added that various environmental conditions can influence how long the virus survives on various surfaces. Those conditions can include relative humidity, fomite material, and air temperature [6].

We weren't necessarily able to take all factors into consideration since data can be limited online, especially at the county level. However, we did work very hard to build a dataset with as many relevant attributes as possible. The data we found primarily came in one of two forms. Either as a dataset such as the list of all the counties in the United State or data spread online such as the percentage of smokers in a specific county. For the latter case, we used different web scrapers based on Python scripts and the BeautifulSoup open library, in order to gather the data and store it in a csv file for later use. The pre-built databases however were not easily appended to our master dataset. A lot of preprocessing and cleaning was required to integrate all of the data from the multiple sources. For example, with all of the weather data, we wanted to look specifically at the weather data from the days since the first infection in each individual county. This required us to first compute the days since the first infection, which was a process in and of itself. Then, we had to parse a past weather dataset and take the mean of the weather data for only those dates that fall within the range of the infected period. All of the different attributes which came from both datasets, and those that we scraped off the web, had similar processes attached to them, which made building the dataset a challenging task.



```
lines = generate_links('states.txt')
f = open("links.txt", encoding="utf-8")
writer = csv.writer(csvfile)
l = 0
fin_rows = []
for link in f:
    state_url = link
    driver = webdriver.Chrome("/Users/ahmedboutar/Documents/chromedriver")
    driver.get(state_url)
    driver.implicitly_wait(300)
    content = driver.page_source
    soup = BeautifulSoup(content, 'html.parser')
    for row in rows:
        curr_state = str(lines[l].lower().strip())
        row_state = str(row[2].lower().strip())
        row_county = str(row[1]).lower().strip()
        perc_table = soup.find('table', attrs={'class': 'measure-data-table sticky-enabled'})
        if perc_table is None:
            print("Table not found for ", curr_state, " state!!")
            print("====================================")
            continue
        if(row_state==curr_state):
            for tr in perc_table.tbody.findAll('tr'):
                county_name = tr.find('td', class_='name')
                percentage = tr.find('td', class_='raw_value')
                if(row_county==county_name.text.lower().strip()):
                    row[10] = percentage.text
                    fin_rows.append(row)
    l = l + 1
    driver.close()
```

```
filename = "allCountyData1.csv"
fields = []
rows = []
# reading csv file
with open(filename, 'r') as csvfile:
    # creating a csv reader object
    csvreader = csv.reader(csvfile)
    # extracting each data row one by one
    for row in csvreader:
        rows.append(row)
    # get total number of rows
    print("Total no. of rows: %d" % (csvreader.line_num))

filename = "conreport2019.csv"
rows_fin = []
with open(filename, 'r') as csvfile1:
    csvreader = csv.reader(csvfile1)
    r = 1
    for row in csvreader:
        for r in rows:
            query = str(r[1]) + " County, " + str(r[3])
            if(query==str(row[1])):
                new_row = r + row
                rows_fin.append(new_row)
```

Figure 3. Snippet of the web scraper code

Web scraping was done in two different ways: the data representing is either scraped from a single page or from different pages. We either determined a relationship between the different URLs or triggered a JavaScript function such as the click() function if the query is only triggered by a user click. For instance, in order to get the percentage of smokers by county, we generated a list of links to the dataset of each county. Then, based on the state whose database is displayed, we updated the percentage of smokers in our own dataset by extracting that piece of data from the web. Some Python scripts were used for transferring data as well as joining datasets together.

2.3 Preprocessing

Initially, the dataset was very large and had a lot of noise. The dataset started with 3,007 counties, which not only makes the processing hard, but also decreases the accuracy of our result considering the noise in the data. We decided to only take into account counties that registered over 50 cases and that reported their first case at least 25 days ago. We think that counties with less than 50 cases won't have a significant impact on our results since we are barely noticing any spread. Using the 25 days mark allowed us to get more significant data as the data collected would mainly be one before the mandatory quarantine started in the United States. Besides, this allows for the data to be temporally aligned.

Some parts of the data we found was incomplete. Some tuples had no recorded values for some

of the attributes, especially those representing the air quality. Since we didn't want to ignore the tuples as we have already significantly decreased the size of the dataset, we filled the missing values with the attribute mean of a state. For instance, we determine the missing value of an attribute A by taking the mean of all the values corresponding to the same geographic location (. i.e. the state to which the county belongs to) of that specific attribute. We also combined computer and human inspection in order to clean the noisy data that didn't fit the general tendency of values in a given state.

3. Model

3.1 Overview

We used a multivariable linear regression learning model to find attribute importance related to rate of spread, which was our pivot attribute. We built the model in Ruby, making use of Ruby's Matrix data structure. Our model used the normal equation to learn attribute weights. Our model reads input from a CSV file and dynamically adapts to the number of independent attributes. Our model also dynamically handles the location of the pivot attribute within each row.

3.2 Accuracy Evaluation

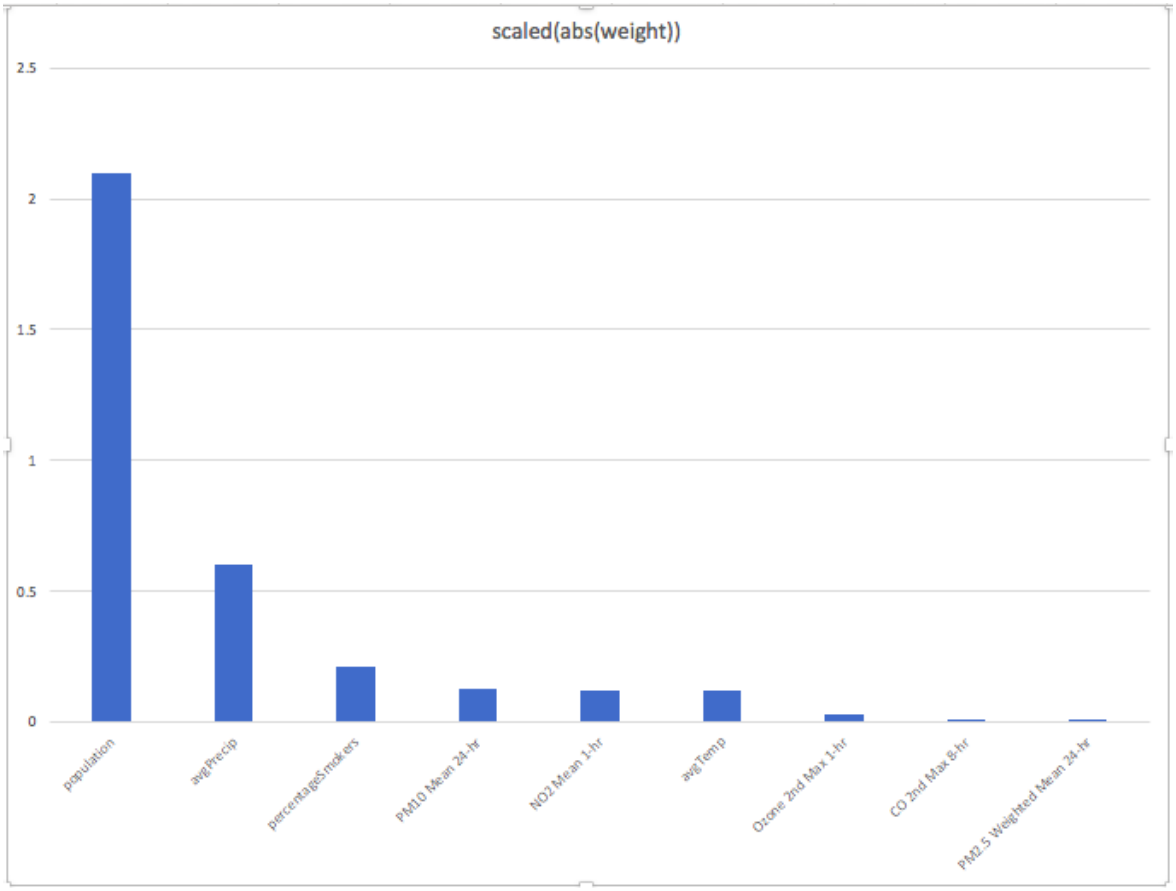
Our model uses mean squared error across all of the independent attributes to assess accuracy.

$$\text{MSE} = \frac{1}{n}\sum_{k=0}^{n}(y_k - \gamma_k)^2$$

- n: number of data points
- y_k: represents observed values
- γ_k: represents predicted value

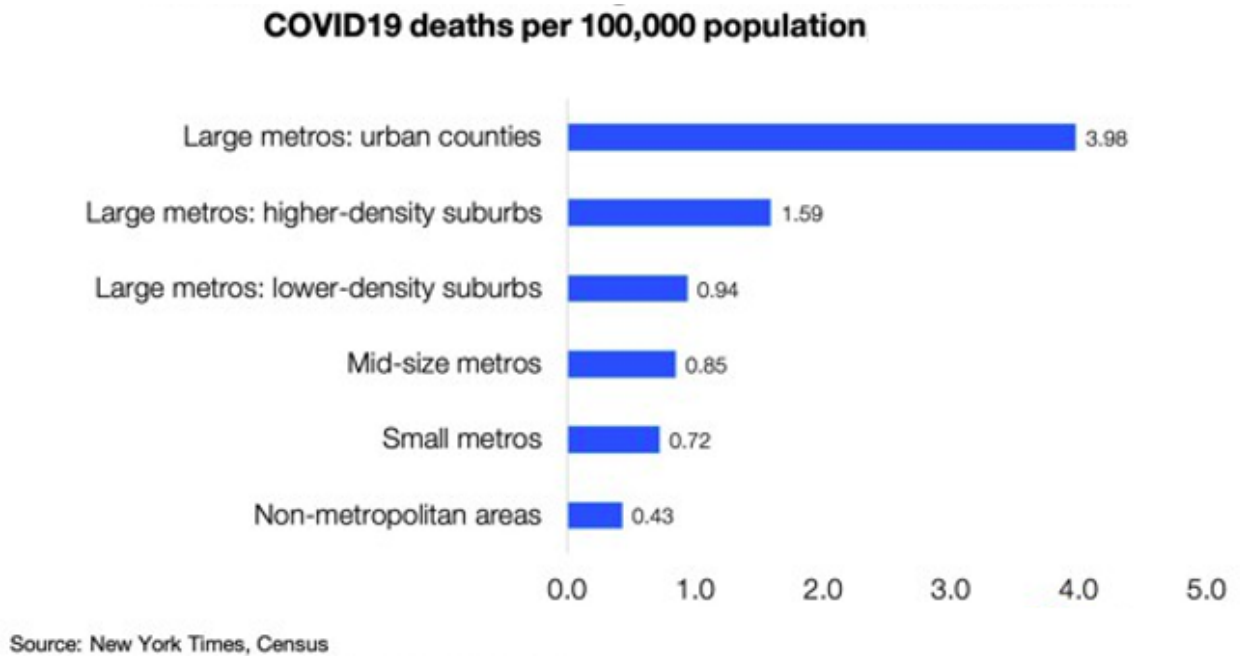
4. Analysis

4.1 Attribute Importance Analysis



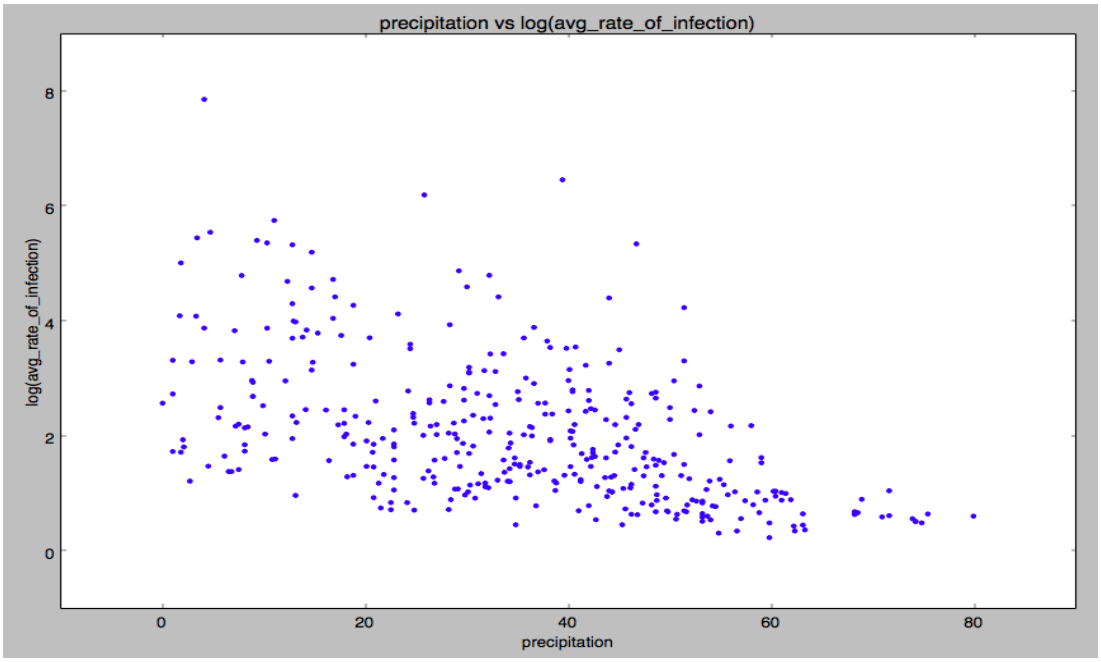
4.1.1 Population Attribute

As we can see from this graph of attribute importance, population was by far the attribute identified as the most significant by the model. The model assigned a massive weight to population relative to the rest of the attributes. Further research into other studies on the relation of population and spread of COVID-19 confirms this finding. Here is a chart from a study done by the New York Times, which shows the per capita deaths in different classifications of US Counties [7]. Those counties which are classified as more “population dense” have much higher per-capita rates of death than less dense ones.



4.1.2 Precipitation Attribute

As can be seen in the attribute importance figure above, the other attribute that was assigned some importance by the model was the average precipitation attribute. To investigate this relationship further, we created a scatter plot of rate of spread and average precipitation data:



This plot indicates a potential negative correlation between precipitation and average rate of infection. One theory for this relationship is that perhaps in places where it rained more frequently during the beginning of the spread, people stayed inside more, which may have helped limit the spread of the virus.

4.2 Model Accuracy

We assessed our model's accuracy using mean squared error across all of the independent attributes. The model was fairly accurate, but it was mostly due to the strong correlation of population and rate of spread. When we ran the model on a modified dataset without this attribute, it was highly inaccurate.

4.3 Findings

Our model's most concrete finding is that there is a strong correlation between population and rate of spread. This finding is confirmed by other studies. Our model also found that there may be a negative relationship between average precipitation and rate of spread. We believe that this may be due to the fact that when it rains, people tend to stay inside, which may help stop the spread of the virus.

Perhaps more important than the findings rooted in attributes with high importance, is the findings, or lack thereof rooted in attributes with low importance. For us, the most surprising non-correlation was that of temperature and rate of spread. The model placed almost no weight on the average temperature attribute. This was surprising to us, because there were some studies earlier in the COVID-19 crisis which seemed to suggest that the virus had smaller rates of spread in places with warmer climates. Our model does not back up this idea.

Our final finding was a lack of correlation between high median age and rate of spread. When we plotted median age and rate of death, there was some visible correlation, but with rate of spread there was not. This supports the idea that while older people are much more likely to die from COVID-19, it spreads similarly regardless of age.

5. Conclusion and Future Work

According to our analysis so far, we have found that the spread of the SARS-CoV-2 virus is correlated to some factors. In fact, our model shows the existence of a positive correlation between the population density and the spread of the virus as well as a negative correlation between the precipitation rate and the spread, which we attributed to a lack of movement among the population on rainy days.

Based on the factors that we were able to gather so far, our model was able to show the correlation between these factors and the rate of spread of the virus and categorize it as positive or negative. However, we believe that more data can be gathered about numerous other factors that we may discover as research on this virus continues. Our model is based on quantitative factors, and we think that incorporating qualitative data may lead to some interesting findings.

Bibliography

[1] "Coronavirus disease 2019 (COVID-19) Situation Report –94". World Health Organization, April 23, 2020, [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.pdf?sfvrsn=b8304bf0_2#:~:text=Retrospective%20investigations%20by%20Chinese%20authorities,%2C%20some%20did%20not.

[2] "Situation Summary". Center For Disease Control and Prevention. Updated April 19, 2020, [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/summary.html

[3] Dantas, R., De Campos, P., Rossi, I., & Ribas, R. (2020). Implications of social distancing in Brazil in the pandemic period of COVID-19. Infection Control & Hospital Epidemiology, 1-5. doi:10.1017/ice.2020.210. [Online]. Available: https://www.cambridge.org/core/journals/infection-control-and-hospital-epidemiology/article/implications-of-social-distancing-in-brazil-in-the-pandemic-period-of-covid19/DF33AB03455FBEFEB9EF88EB95E789B4

[4] Sean Wei Xiang Ong, Barnaby Edward Young, Yee-Sin Leo, David Chien Lye, Association of higher body mass index (BMI) with severe coronavirus disease 2019 (COVID-19) in younger patients, Clinical Infectious Diseases, , ciaa548, Available: https://doi.org/10.1093/cid/ciaa548

[5] D.Levine, “Does Smoking and Vaping Make Coronavirus Worse?”, US News & World Report, March 31, 2020. [Online]. Available: https://health.usnews.com/conditions/articles/smoking-vaping-coronavirus

[6] A.Antrim, “Environmental Engineers Identify Factors Affecting COVID-19 Transmission”, Contemporary Clinic, March 30, 2020. [Online]. Available: https://contemporaryclinic.pharmacytimes.com/news-views/environmental-engineers-identify-factors-affecting-covid-19-transmission

[7] CityLab and University of Toronto’s School of Cities and Rotman School of Management, “What We Know About Density and the Spread of Coronavirus,” CityLab, 17-Apr-2020. [Online]. Available: https://www.citylab.com/equity/2020/04/coronavirus-spread-map-city-urban-density-suburbs-rural-data/609394/. [Accessed: 09-May-2020].