Decoding Covid-19's Correlation to Environmental and Cultural Factors

Nikita Patel
CS613 Machine Learning
nrp62@drexel.edu

***Abstract*—The Coronavirus (Covid-19) disease, similar to SARS-Cov-2, is a betacoronavirus having its origins in Wuhan, China in December 2019, which quickly spread from region to region. In a few short months, cases were popping up in nearly every country across the globe. This virus, as most viruses do, exhibits exponential trends in spread and primarily affects adults and people who have severe underlying medical conditions, with patients experiencing severe illness and death. Covid-19's rate of growth has been a major concern since the beginning of the pandemic due to the limitations of hospitals, personal protective equipment, and ventilator equipment. Understanding the features contributing to its spread have become a topic of much research since its origin. Numerous researchers have performed studies evaluating the contributions of the different features such as population of a region, race, and even environmental features like temperature. This study evaluated several environmental, cultural, and census related features, and was not able to find any strong correlation between any of the above features and the number of new coronavirus cases in a region per day.**

Introduction

The Covid-19 pandemic has been a unique experience on a global scale, with no other virus in the modern world having had a similar impact. In the last year, there have been 69.2 million cases and 1.5 million deaths worldwide, with 15.8 million cases and nearly 300,000 deaths just in the United States of America [1]. Around February of 2020, when the world outside of China was first beginning to fear or was experiencing the initial spread of the virus, much of its characteristics were unknown and very little data was available to study. Within the past year, there have been numerous sources collecting relevant infection and death data, as well as several publications on what researchers believe to be the key factors in the virus's ability to spread so quickly.

The virus is transmitted through close contact with an infected individual, and therefore it was suspected that areas with large populations densities would be impacted on a larger scale than more rural locations. However, it has become clear through time and data that population density is not the only factor that should be considered. Several studies have evaluated the relationship between warmer climates with higher humidity levels in comparison with regions with colder climates and the characteristics of their spread. While many were able to show correlation, none were able to show a direct relationship. It is my hypothesis that the spread of the coronavirus is correlated not only to weather characteristics, but also to the physical geography of a region, its ability to control its borders, the implemented government guidelines, as well as the culture of the individuals living in the region. Many of the studies collected were on a global level, which is good to show world trends, but for a region to gain actionable intel, isolated studies must be conducted. Although determining a correlation between temperature and the spread of the disease will not contribute to the virus's elimination, it can provide insight to allow populations to operate with caution and within boundaries during certain weather periods.

Additionally, it can in-turn allow governments to make more informed decisions on what regulations to put in place and help secure the financial futures of their citizens. This paper evaluates weather, population, and cultural features from January to November 2020 across all 50 states to see which seem to show the largest correlation between the number of new cases in a day. The data used was collected for the largest county in each state, and was evaluated using a variety of regressor models including Linear Regression, Support Vector Machines, K-Nearest Neighbors, Decision Trees and Random Forest.

Related Works

As mentioned, several studies have attempted to evaluate the relationship between the spread of Covid-19 and weather factors such as temperature. Zohair Malki et al [2] evaluated the relationship of temperature and humidity and the spread of the disease across Italy on March 16, 2020 and 17th, taking weather data from one day and correlating it to the number of cases on the next. The authors were able to indicate that both features played somewhat significant roles in the mortality rates of a region, and were thus able to loosely indicate that as temperatures in a region went up, infection rates went down. While this study showed a correlation worthy of further testing, it was somewhat limited in its scope. It only evaluated temperature features on a single day and used weather data from the day prior in attempts to draw a correlation. In order to show or prove a more significant relationship, this paper will be using a range of dates across 2020. The referenced study was conducted in June of 2020, and since, twice as much data has become available simply due to time passed. In the past 6 months, the United States, in addition to most countries in the northern hemisphere, have experienced a second spike in coronavirus cases, even causing many countries to go into their second-wave shutdown. The waves indicate highs in lows in the spreading of the virus that correlate to the colder and

warmer months respectively and therefore hint at the viruses' seasonality aligning with the findings of the authors.

J. Wang et al [3] performed a similar study, largely using the same set of features, but instead took a more global approach. They were inspired by observations made in the beginning months of the pandemic and noticed that cities with climates similar to Wuhan were being more greatly impacted. In the beginning evaluations of their data, the authors noticed an inverse relationship for both temperature and humidity features. But upon applying filters and cleaning their data, they concluded that no clear relationship existed.

Several other papers exploring the same topic seem to be on both sides of the fence, some seeing a correlation, although not strong, and others not seeing one at all. This could be due to the scope of the studies. While some looked at isolated regions, others took more global perspectives but did not study every country. From my perspective, the existence of a variety of results depending on the training set indicate that there are several features that contribute to the spread, and it cannot be captured simply by a feature like temperature or humidity.

Many parallels have also been drawn between influenza and Covid-19 in regard to their seasonality. Influenza is a virus that has been studied for decades, and if any parallels or links can be made with the coronavirus, it would mean leaps and bounds for research. Anice C. Lowen, John Steel in their study Roles of Humidity and Temperature in Shaping Influenza Seasonality [4] conduct transmission experiments on guinea pigs. They were able to "[identify] absolute humidity and temperature as climatic predictors of influenza epidemics in temperate regions of the world". They were unable to identify the exact mechanisms by which these environmental factors alter transmission, but the relationship was described as robust. A similarly conducted controlled experiment of coronavirus transmission could potentially exhibit a similar result indicating that a variety of external and unaccounted for factors are affecting existing research.

METHODOLOGY

The first step was to collect all the potential data that could be considered relevant to the spread of the virus across census, weather as well as cultural perspectives. There were a total of eleven features evaluated in this study in the following categories. Weather data was collected primarily from the National Oceanic and Atmospheric Administration (NOAA) [5]. Daily minimum and maximum temperature readings as well as average daily wind speed from the most populated cities in each US state was manually requested between the dates of January 1st and November 22nd. Temperature on a given day bears no correlation to the number of new cases on that very day. Studies indicate that individuals will start to see symptoms anywhere between 2-14 after infection so for each day in each location, the number of new Covid-19 cases were correlated with the average max and min temperatures and average daily wind speeds for the preceding two weeks.

Census features and population information of various types were collected for different reasons and based on different hypotheses. To test the effects of Covid-19 on cities in relation to rural areas, the population density was used as a feature. Data on the percentage of people below the poverty line in a county was also used considering poorer communities have been seen to be afflicted by the pandemic disproportionately. Age was also a factor believed to be of importance. This was not only because it is believed that older populations are more likely to be affected, but also because younger populations are more likely to be asymptomatic. This leads to unknowingly spreading the virus and potentially causing higher rates of infection. Therefore, data on the percentage of the population below 5 years old, between the ages of 5 and 18, as well as the percentage of the population above the age of 65 were incorporated into this study. All this data was collected from the 2010 census [6].

An attempt was made to incorporate cultural aspects into the studied features, which included the prevalence of mask usage per US county [7], as well as the travel frequency and distance of the population of counties collected from the Bureau of Transportation Statistics[8]. Specifically, I gathered data on the percentage of people that wear their masks more than what was defined as regularly and the percentage of the population that stays at home. These are key features encompassing the degree to which populations are following CDC guidelines and are indicators on the perspectives of the population on the severity of the pandemic.

Lastly, data related to the number of infected individuals was leveraged from New York Times [7] databases. From this data, the number of new cases per day, per county, as well as days since the first Covid case in a US were derived. The parameter of 'days since the first case' was incorporated in effort to factor in time. Intuitively it can be hypothesized that for a disease that exhibits an exponential growth, as time passes, the rate of growth will continue to increase.

A variety of other features were also collected including testing prevalence of a location, but it was ultimately decided that correlation between such features and new cases per day would be a misleading correlation since a lack of testing can show areas as being less affected than they truly are and vice versa.

Upon the collection and cleaning of the data, various regression models were chosen to evaluate the selected features. Using inspiration from [2] and their results, Linear Regression, Support Vector Machines, K-Nearest neighbors, Decision Tree Regression, and Random Forest Regressor models were incorporated. Significant consideration was taken in potentially conducting the study using classification models instead. However, classification of certain features such as population density were somewhat vague and arbitrary, and the method did not seem to provide any foreseeable benefits. In fact, since each feature had continuous data, it was concluded that regression would be the most effective technique. Various permutations of features were also tested against the data and their errors were recorded to be evaluated. The following error functions are used: root-mean-square, mean-square, mean absolute error, r2 and mean absolute percentage error, similar to those used

in [2].

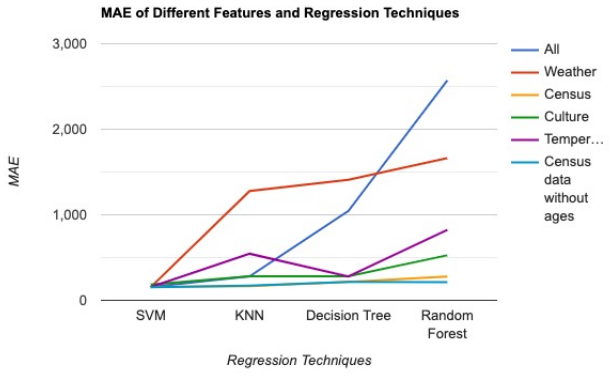
Experiments and Results

The full data set of all eleven features was evaluated all at once by running the data through the variety of regression models. The results can be seen in Table 1.

Each error calculation method produced high values, indicating that there is no correlation between the provided dataset and the number of new cases in a day in a given county. Something interesting is that the K Nearest Neighbors algorithm, giving values of k between 1 and 10, a k of 1 produced the best results. In fact there is a very clear and steep relationship between the values of k and the produced errors. Low k's are typically avoided because they are easily influenced by noise. The preference of a low k by the dataset could be an indication of randomly scattered data and that no clear groupings exist.

The next step was to try and break the eleven features into subsets and evaluate them individually. Subsets of the data for the environmental/weather variables, population/demographic variables as well as cultural variables were created and run through the regression models and their results can be seen in the following tables. Table 2 incorporates the features of average maximum temperature, minimum temperature, and daily wind speed for the preceding two weeks, and the errors show the strength of the correlation. Table 3 evaluates the features of age, population density, and percentage of population in poverty, and Table 4 evaluates mask and stay-at-home features.

In attempts to get error values even lower, subsets were created within the ones created above. There was no significant change in the MAE error for the svm regressor. However, improvements in the other regressors's error values were as seen in Figure 1. As the subsets got smaller, improvements were visible in some regression models, but only slightly. In The mean absolute error seemed to be the lowest error value across the board, however, that does not quite mean anything. Root-mean-squared, means-squared error, and mean absolute error express average model prediction error. It is just the nature of their calculation that makes their values higher or lower, but MAE just seemed the best to show graphically how similar each of the subsets were.



When looking at the evaluation of the dataset of all eleven features (including census, weather and cultural), and looking at the evaluations of the subset, no clear correlation exists for any subset compared to any other subst. Weather features seem to show no more of a significant correlation with the number of new cases a day than demographic/population related features as in line with the findings of [1] and [3].

Conclusion

In line with the 'no free lunch' paradigm, it was important to see which regression algorithms worked best for the scenario at hand, which is why a variety of models were used in testing. In all of the tests done, it is made obvious that the data is not linear in nature due to how poorly the linear regression model performed and also by how well the SVR performed, since it is designed to acknowledge non-linearity. In this implementation, the SVM regression proved to show the strongest results in terms of correlation, it exhibited the lowest error across all used error calculation techniques across almost every test with KNN in a close second.

No direct or strong correlations between the features or subset of features were detected when predicting the number of new cases per day in a given region. In fact, the MAE error for the SVN regression for the full data set and each subset varied minimally except for when assessing the strength of the cultural feature subset, it had a slightly higher error than all others. A reason for this could have been due to the quality of the data since that majority was accumulated through polling, it may have slightly, inaccurate representations of the truth at any given time. There could be many reasons attributing to the fact that no significant correlation was seen overall, one being that none of these features have any correlation with the spread of Covid-19, however, that is believed to be highly unlikely. While many studies could not conclude any strong correlations, the reason for no correlation could potentially be attributed to the variety of the locations where the data was drawn from.

When the data was stripped of the location from where it was collected, it lost a significant amount of context which none of the other provided features could make up. Every state in the US, or at least every region in the US is culturally different. In New York City, people tend to spend their time indoors and generally in close proximity to one another, whereas in a place like Montana the opposite is true. When trying to evaluate why a disease relying on human-to-human contact spreads more in one location, that information is extremely relevant. This goes to prove that there are further cultural aspects associated with the spread of the disease we will have to work hard to understand.

Future Work and Extensions

Like many of the papers on this same topic, the results were far from conclusive. However, there are various areas and ways in which this study could be improved to potentially show a correlation between the data features and the number of new Covid-19 cases in a day.

As time passes and more data on the disease is captured, it would be beneficial to re-conduct the study, potentially with

TABLE I
REGRESSOR MODEL EVALUATION FOR DATASET OF ALL 11 FEATURES

Labels	MAE	MSE	RMSE	R2	MAPE
Linear	76872122.43	5.72506E+19	7566410968	-123118071956	2.86188E+24
SVM	157.5233044	485830981.5	22041.57393	-0.04478844936	3.52281E+18
KNN (N=4)	280.0231339	532439817	23074.65746	-0.1450216062	1.19E+19
Decision Tree	1046.814621	3917421934	62589.31166	-7.429581057	4.39329E+19
Random Forest	2570.657563	21821416948	147720.7397	-45.97011185	1.01856E+20

TABLE II
REGRESSOR MODEL EVALUATION FOR DATASET OF WEATHER RELATED VARIABLES

Labels	MAE	MSE	RMSE	R2	MAPE
Linear	978.6066683	4454321883	66740.70634	-8.256879709	4.03595E+19
SVM	154.8358991	499173679.7	22342.19505	-0.073482166897	2.6847E+18
KNN (N=4)	1279.238279	5604527530	74863.39245	-10.95865579	5.30704E+19
Decision Tree	1408.620296	6641075649	81492.79507	-13.28175516	5.78911E+19
Random Forest	1663.260827	9191007562	95869.74268	-18.76543058	6.75844E+19

TABLE III
REGRESSOR MODEL EVALUATION FOR DATASET OF POPULATION RELATED VARIABLES

Labels	MAE	MSE	RMSE	R2	MAPE
Linear	96033.15336	75773520613803	8704798.712	-162881.0157	3.58173E+21
SVM	155.1958408	513540925.7	22661.44139	-0.1043791932	1.9515E+18
KNN (N=4)	165.0753393	493843323.6	22222.58589	-0.08510494982	3.24036E+18
Decision Tree	216.4188026	473746782.3	21765.72494	-0.01880115689	8.44001E+18
Random Forest	278.6514693	530722164.9	23037.40795	-0.1413211922	1.18072E+19

TABLE IV
REGRESSOR MODEL EVALUATION FOR DATASET OF CULTURE RELATED VARIABLES

Labels	MAE	MSE	RMSE	R2	MAPE
Linear	84572708.47	6.92926E+19	8324218250	-149014654275	3.14857E+24
SVM	182.9267415	465201487.9	21568.53004	-0.0004243443911	6.21629E+18
KNN (N=4)	280.0231339	532439817	23074.65746	-0.1450216062	1.18761E+19
Decision Tree	280.0231339	532439817	23074.65746	-0.1450216062	1.18761E+19
Random Forest	526.5829118	1104188877	33229.33759	-1.374578461	2.30373E+19

the same features, to see if there is any improvement in the results. As more data becomes available, it would also be interesting to do more isolated studies on different regions. For example, a study can be conducted on just the state of New York to see if correlations seem stronger or weaker. Each state in the US has a variety of features and contexts that are difficult to represent, and therefore doing the study state-by-state could potentially yield more conclusive results.

The depth of the existing high level features could also use improvement. Many features contribute to a climate including, but not limited to, humidity, rainfall, sunshine hours, UV index, etc. If a correlation between climate and the growth in cases is to be drawn, a more specific and thorough study needs to be conducted.

Lastly, while extremely difficult, the incorporation of more in-depth cultural values would be very interesting. While things like temperature may have an effect one way or another on a disease's spread, the spread happens through human-to-human transmission. Thus, people are the primary factors helping a disease to grow, as is made obvious through social distancing measures and stay-at-home orders. It would be interesting to capture the percentage of people that spend time outdoors rather than indoors, and which social habits people have for different regions to factor that into the existing study.

REFERENCES

[1] "Coronavirus Cases:" *Worldometer*, www.worldometers.info/coronavirus/?utm_campaign=homeAdUOA%3FSi.

[2] Zohair Malki, El-Sayed Atlam, Aboul Ella Hassanien, Guesh Dagnew, Mostafa A. Elhosseini, Ibrahim Gad. Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches

[3] Jingyuan Wang, Ke Tang, Kai Feng, Xin Lin, Weifeng Lv, Kun Chen, Fei Wang. High Temperature and High Humidity Reduce the Transmission of COVID-19

[4] Anice C. Lowen, John Steel. Roles of Humidity and Temperature in Shaping Influenza Seasonality. *Journal of Virology Jun 2014*, 88 (14) 7692-7695

[5] National Centers for Environmental Information (NCEI). "Climate Data Online Search." *Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC)*, www.ncdc.noaa.gov/cdo-web/search.

[6] "Census Population Data." github.com/ykzeng/covid-19/tree/master/data.

[7] "Coronavirus (Covid-19) Data in the United States." *New York Times*, github.com/nytimes/covid-19-data.
[8] "Daily Travel during the COVID-19 Public Health Emergency." *Bureau of Transportation Statistics*, www.bts.dot.gov/daily-travel.