Math 189Z: Final Project

Rachael Soh

15 May 2020

1 Motivation

The objective of this study is to explore the relationship between factors of weather and the transmission of Covid-19. Over the past few weeks, many news sources and medical experts have begun contemplating the possibility of high temperature and humidity slowing down the spread of the Covid-19 [5] , some even hoping that summer will bring this pandemic to an end [6]. This possibility is derived from the fact that both SARS and influenza have shown similar patterns in the past, and considering that Covid-19 transmits very similarly to these other viruses, this behavior should be observed as well given the same high temperature and humidity environment. [8]

2 Methods

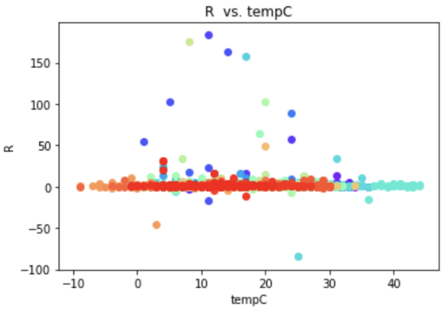
The main factors of weather our team will be studying are: temperature, humidity, wind and precipitation. Furthermore, the analysis will focus on analyzing this relationship in the U.S as we felt that it was the most relevant to us and has a diverse mix of climates which could yield interesting results. The time period will begin on January 20, the first reported case in the U.S. up to the most current date. We used the WorldWeatherOnline [7] historical weather data API wrapper `wwo_hist` to pull historical data of specific cities or regions and the Covid-19 cases and deaths count from the John Hopkins Github page.[4]

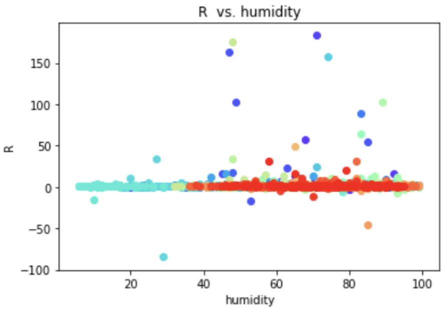
We will mostly be running linear regression of the basic reproductive values, R, against these four different factors. We began by running univariate regression of each weather factor against the R value and then ran a multivariate regression with all the weather factors with and without the interaction terms[1]. This analysis was run at a state level because we were unable to find an accurate and simple way to get city level weather data.

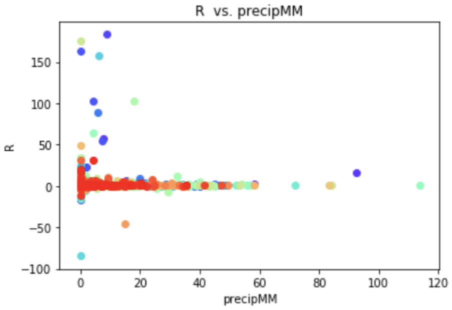
Once we ran the multivariate regression of only weather elements, we found indication of collinearity within the model so we decided to drop the precipitation variable in an attempt to reduce this. We also added ”Population density” as another explanatory variable to create a better model [2]. Finally, we tried to find which covariates contributed significantly to the R values of each state.

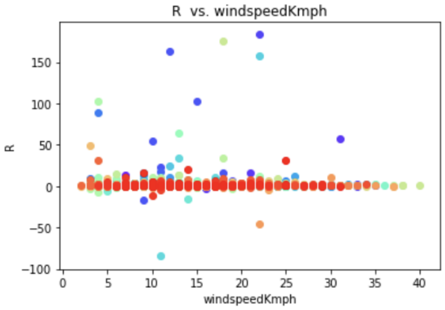
3 Results

When the R values were plotted against the different weather variables, we found that there was no obvious pattern.







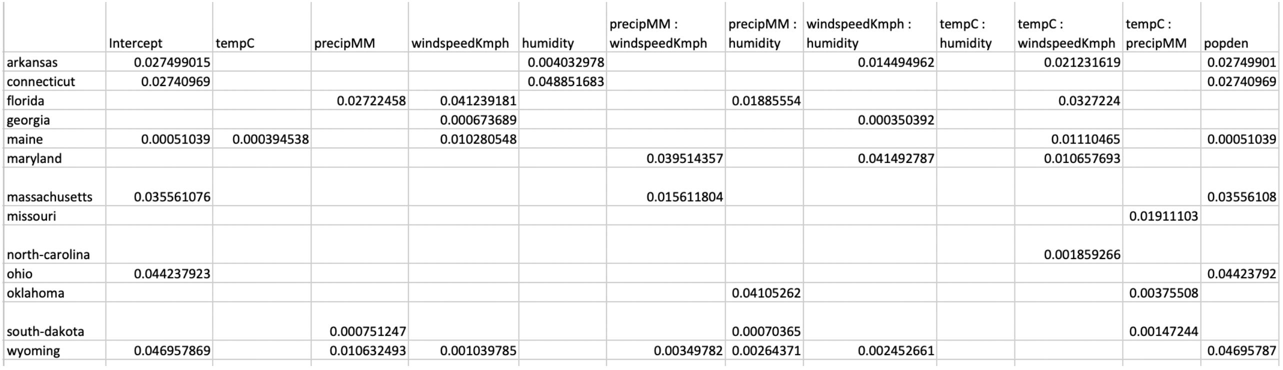


As such, we ran the univariate regressions and found that each weather element was significant in 8 different states, all of which were different in each factor and only Nebraska appeared more than once.

Weather variable	State	p-value
Temperature	Maryland	0.015248
	Missouri	0.010844
Humidity	Indiana	0.004938
	Nebraska	0.037461
Rain	Arkansas	0.033296
	Nebraska	0.000570
Wind	Arizona	0.017244
	Illinois	0.048262
	North Carolina	0.048263

We then ran the multivariate regression with and without the interaction terms and the additional explanatory variable population density. However, we found that our results with the interaction terms explained the data better as it had a higher R-squared value. Additionally, the condition number was extremely high, so we tried dropping the precipitation variable in an attempt to reduce this condition number because we felt that precipitation and humidity would be have collinearity. Despite dropping precipitation, the condition number remained extremely high and now our R-squared value was much lower, so this wasn't very promising and decided against removing it.

Once we found the summaries of all the multivariate regressions of each state, we found all the occasions in which a variable was significant and then took a closer look at their relationships. Unfortunately, none of the coefficients were consistent in all the states so we cannot draw any conclusions from our data analysis. Ex. wind speed was associated with a positive change in the R value in Georgia but a negative one in Florida.



	Intercept	tempC	precipMM	windspeedKmph	humidity	precipMM : windspeedKmph	precipMM : humidity	windspeedKmph : humidity	tempC : humidity	tempC : windspeedKmph	tempC : precipMM	popden
arkansas	0.027499015				0.004032978			0.014494962		0.021231619		0.02749901
connecticut	0.02740969				0.048851683							0.02740969
florida			0.02722458	0.041239181			0.01885554			0.0327224		
georgia				0.000673689				0.000350392				
maine	0.00051039	0.000394538		0.010280548						0.01110465		0.00051039
maryland						0.039514357		0.041492787		0.010657693		
massachusetts	0.035561076					0.015611804						0.03556108
missouri											0.01911103	
north-carolina										0.001859266		
ohio	0.044237923											0.04423792
oklahoma							0.04105262				0.00375508	
south-dakota			0.000751247				0.00070365				0.00147244	
wyoming	0.046957869		0.010632493	0.001039785		0.00349782	0.00264371	0.002452661				0.04695787

4 Discussion

After our data analysis, we still have not found any significant relationship between weather and the R value (infection rate) of the Covid-19. At most, our weather variables with the interaction terms only have an explain about 0.2 of our variation in the R-values (R-squared around 0.2). It seems that we are missing some very important variables in our data, and surprisingly population density did not contribute much to the multivariate model's R-squared value. We think that this issue, as mentioned in our progress report, is largely due to the fact that we are running weather data on a state level when we should be doing it on a smaller city-scale level to give more accurate results. Ex. San Francisco and Los Angeles are both in California but have different climates since California is so big. However, this stems from an issue we have been struggling with since the beginning: an simple source we can pull accurate weather data from. As of now, `|wwo-hist|`works well for state-scale weather data, but there were some cities we could not pull from it. We also need to find out the source of collinearity in our model. We initially suspected that rain and humidity was the cause of this collinearity but once we tried dropping rain, the condition number still remained extremely high which means that there was another source of collinearity in the model. We think that we should run PCA to get independent combinations of variables and after performing your regression you can use the eigenvectors to determine the important variables.

Another problem that I suspect is giving us inaccurate results is that our data is a time series, so instead of running OLS on the data I would like to run regression which works on panel data.[3] I think that running a fixed effect model could solve this problem, as to remove heterogeneity bias, and will be the next step I am taking to finding a better model for this data. Additionally, I am currently running OLS on each state, but I may pool all the data together. The main reason as to why I did it on each state is because `wwo-hist` downloads a csv for each state and so when I load the csv in, it produces one data frame per state. Another reason is that if I pooled all the data, I could lose state specific demographics like ”population density”. Pooling the data is still something we are considering and would need further research and study before doing.

Although we were not able to conclude anything from our data analysis, we still felt that these weather variables still had an interesting relationship with the R-value in each state. It was interesting that some of the worst hit states like New York and California did not have a significant hit in any of these weather variables but it made sense as there must be some other covariate, which we did not account for, that explains the variance in the R values. Perhaps finding these more significant covariates could help us get a better model so we will continue exploring the options. In the proposal we mentioned that we hope to find a significant relationship or model to predict possible R values given historical weather data, but as of now we are still unable to do this. We also had the idea to build a model or jupyter notebook that can monitor the weather and numbers of cases, deaths or R value in a city, state or country in the future but have not progressed on this idea as we continue to struggle in finding good weather data.

References

[1] Interactions in regression.

[2] Regression analysis with control variables.

[3] Introduction to regression models for panel data analysis, 2011.

[4] CSSEGISandData. Cssegisanddata/covid-19, 03 2020.

[5] Richard Gray. Will warm weather really kill off covid-19?

[6] Stephanie Pappas-Live Science Contributor 19 March 2020. Will covid-19 die down in summer? new tests could help answer that.

[7] Ekapope Viriyakovithya. wwo-hist: This package is used to retrieve and transform historical weather data from www.worldweatheronline.com into pandas dataframe and csv.

[8] Jingyuan Wang, Ke Tang, Kai Feng, and Weifeng Lv. High temperature and high humidity reduce the transmission of covid-19. *SSRN Electronic Journal*, 2020.