Kristine Dinh

STAT 673: Journal Article Assignment

December 9, 2020

Time-Series Analysis for the Association of COVID-19 and Climate in Mainland China

**Executive Summary:** COVID-19 is a deathly illness that caused by a coronavirus started in late 2019. Many doctors and scientists want to find the variables that effect the infection of this virus. Knowing the variables that effect the infection of coronavirus, everyone in the world can be more aware to avoid getting infected. One of the most obvious factors is climate change over time. An analysis by Qi et al. (2020) was analyzing a time series relationship between COVID-19 and average temperature and relative humidity scores in Mainland China. The authors include a study of COVID-19 cases in 31 Chinese provinces over the period of two months in early 2020. They found a negative relationship between climate factors and COVID-19 cases using a generalized additive model. In other words, if the temperature is higher, on average, the number of COVID-19 cases will decrease. In addition, if the values of relative humidity are higher, then the number of COVID-19 cases will also decrease. Some significant finding that Qi et al. mentioned were a significant interaction between COVID-19 and climate variables have a 95% confident interval of 0.004 to 0.07 in Hubei, one of China’s provinces. This indicates that, on average, as average temperature in this province increases by 1 degree Celsius, the daily confirmed COVID-19 cases will decrease by 36% to 57% while average relative humidity is between 67% to 85.5%. In addition, on average, as average relative humidity increases by 1%, the daily COVID-19 confirmed cases will decrease by 11% to 22% while average temperature is between 5.04 and 8.2 degree Celsius. Qi et al. also mentioned that the number of changes is vary based on province in Mainland China.

**Introduction:** According to the World Health Organization (WHO) COVID-19 has became a pandemic in the late 2019 and beginning of 2020. The number of confirmed cases has been rising exponentially. Most countries must stop operating indoor business, gathering, and unnecessary traveling. Since this is such a significant issue that would cause the economy to be in regression, scientist and research would come up with many questions in favor of this issue. One of the most obvious questions would be exploring the relationship between COVID-19 and meteorological factors including temperature and relative humidity overtime. With the answer to this question, health care professional would expect to have more patients in certain season compared to others. This way, they can prepare accordingly for the season to have enough supplies for the patients. In addition, normal citizens can have a better understand of which time of the year the coronavirus would be more infected. This way, they can stay home when they need to. The purpose of Qi et al.’s analysis is to prove the association of environmental factors and COVID-19, scientifically. With the historical data of confirmed cases in Mainland China, the authors will signify the correlation between temperature and relative humidity and the infection of coronavirus.

**Methodology:** The COVID-19 confirmed cases data were collected from the National Health Commission of People’s Republic of China from December 1, 2019 to February 11, 2020. The data date ranges are depending on each province. The climate change data are from the Weather Underground. This dataset includes meteorological variables including the average temperature and average relative humidity of all provinces in China. Using the two data sets given above, a generalized additive model (GAM) will be used for the model development process. Qi et al. consider a model with 14-day exponential moving average function with lag effects of the interaction between temperature and relative humidity. The count of confirmed COVID-19 cases was also transformed using log-transformation to have a more stationary time series process and smaller values to work with. After the model development process, the authors also validate the model using the same time series model from different provinces including Wuhan and Hubei. All analysis in their paper were done in R with version 3.5.3.

**Results/Discussion:** Before building the model, it is important to assess the relationship between the response variable and each of the predictors. Climate changing over time can be a huge factor in changing the count of COVID-19 confirmed cases. It is found that the daily count increases dramatically on January 20, 2020 in Hubei province when the temperature slightly falling from December 2019 to January 2020. In addition, the percentage of average relative humidity is between 42.17% to 96.92%. Other provinces have an average humidity rate of larger than 88% on January 23, 2020 and February 8, 2020. In other words, the number of cases generally increases as the average temperature and average relative humidity decreases. The count of COVID-19 confirmed cases by each climate factor are visualized in Figure 1. As one can see in Figure 1a, the log-transformation of the daily confirmed cases is decreasing as the relative humidity percentage from the 25th, 50th and 75th percentile increases. In addition, the number log-transformed daily confirmed cases also decrease as the average temperature in the 25th, 50th, and 75th percentile increases. This suggests that these two variables can contribute significant factors to the fluctuation of COVID-19 confirmed cases.

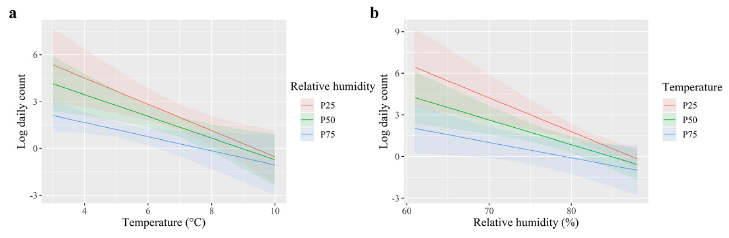
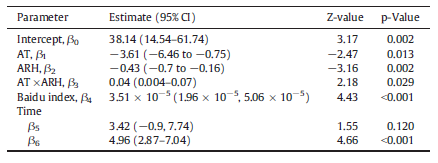


Figure 1. The correlation between log-transformed daily confirmed cases vs climate factors including average temperature and average relative humidity by percentiles.

Now that we have explored the variables and found two significant variables that could impact the daily cases, we can proceed to the model development process. Qi et al. used a GAM model to demonstrate the relationship between target variable and predictors. Table 1 shows the summary regression of the GAM model. This indicates that all variables in the model are significant, including the interaction between average temperature and average relative humidity percentage. Furthermore, the confident interval for all coefficient are clearly saying that the coefficient estimates are significant since there is no zero included in the interval, except one coefficient for time series component.

Table 1. Summary regression of the GAM model for Hubei province. The columns represent parameter of the feature, 95% confident interval of coefficient estimates, test statistics, and significant p-values, respectively from left to right. Looking at the model summary, the p-value for all variable are very small, except for one of the time coefficients.



Now that the model has been developed, some of the coefficients for the GAM will be interpreted. The correlation between daily COVID-19 confirmed cases and climate factor have been found to be negative. Looking at Table 1, the interaction between average temperature and average relative humidity percentage have a coefficient of 0.04 and a 95% confident interval of 0.004 and 0.07. This indicates that if there is an increase in average temperature, the effect of average relative humidity on the daily confirmed cases will decrease at Hubei province. Qi et al. believes that temperature changes could make people’s nasal mucosa prone to small ruptures. Therefore, this could create opportunities for the coronavirus to enter the body easier.

In terms of average temperature, Table 1 shows the relationship between temperature and daily COVID-19 confirmed cases to be negative with an estimated coefficient of -3.61 and a 95% confident interval between -6.46 to -0.75. This indicated that as temperature increases by 1 degree Celsius, the number of confirmed cases will decrease on an average of 4 cases. In addition, average relative humidity percentage also found to have a negative relationship with daily COVID-19 confirmed cases. This indicates that as the percentage of relative humidity increases by 1%, the number of confirmed cases will decrease, on average, by 1 case. Moreover, the time series component of this model is also significant. As seen in Table 1, and are the regression coefficient of natural splines of time with two degrees of freedom. These two coefficient estimates have a positive relationship with confirmed cases. Therefore, this indicates that as time passed by, the number of daily cases would increase.

After the models is finished building, we need to validate the model. In the article by Qi et al., the model is validated by providing the predictors for province Wuhan and Hubei in China using the same GAM. They authors want to see if the model could predict the number of confirmed cases over time for each province. However, the predictions were biased based on time series. Since this model was only built using data from two months period from December 2020 to February 2020. However, the results came out to be like the authors expected. Predicted number of cases in province Wuhan and Hubei seems to be decreases as average temperature and average relative humidity percentage increases. Therefore, this create a negative relationship between target variable and the meteorological factors, while the interaction between average temperature and average humidity percentage is positive.

Despite the biased prediction of the model while outside the range, the model is still good to represent the descriptive relationship of daily COVID-19 confirmed cases and meteorological factors including average temperature and average relative humidity percentage. The final model is presented as the following equation:

,

where is the expected values of daily confirmed cases, is the intercept, is the effect of moving average of average temperature, is the effect of moving average of average relative humidity percentage, is the interaction between average temperature and average relative humidity percentage, is the effect of moving average of Baidu index from the weather data, and and are the coefficient of the splines of time. The estimated values of beta coefficient can be found in Table 1.

**Conclusions**

Since COVID-19 became the center topic in 2020 and a pandemic in the world, scientists are looking for solution to help decrease the daily confirmed cases. Two automatic factors that came into mind are meteorological and time series factors that could affect the daily confirmed cases. Qi et al. have proved in their analysis that there is a negative correlation between the number of COVID-19 confirmed cases and meteorological factors including average daily temperature and average daily relative humidity percentage of each province, especially Hubei. This indicates that we should expect less cases in winter season and more COVID-19 cases in summer season. With this information, health care professionals would have more idea of when to expect more patients at the hospital to prepare enough supplies to treat everyone. In addition, normal citizens would also have a better idea of staying home when the summer comes to avoid getting infected. On the other hand, the interaction between average temperature and average relative humidity percentage was found to be positive. This means that the effect of average relative humidity on expected cases would decreases as the average temperature increases. In addition, time also have an important effect in this model. It is proven that time have a positive effect on the number of cases. Therefore, we would expect the daily confirmed cases to increase as time passes by.

There are limitations to this analysis. Qi et al. mentions that there might be more risk variables that would contribute to rise the number of confirmed cases including social economic status. They also discuss how the analysis is studied in a short period of time and the weather data was only collected from the capital city of each province. This leads to an inaccuracy of other cities in the province since each province have a very large area in China. In addition, this study is only been done on the months on December 2019 to February 2020 specifically in Hubei province. If we want to project the model to data from different time or different province, the information might be unstable. To further study this project, the authors would study the meteorological factors on COVID-19 in depth. In addition, Qi et al. also mentioned that there is a spatial and temporal component to be explore further.

**Bibliography**

Qi, Hongchao, et al. “COVID-19 Transmission in Mainland China Is Associated with Temperature and Humidity: A Time-Series Analysis.” *Science of The Total Environment*, vol. 728, 2020, p. 138778., doi:10.1016/j.scitotenv.2020.138778.