The Correlation Between Temperature And Transmission Of Coronavirus

2034168

I. INTRODUCTION

This data analysis uses two datasets[3][4] from Kaggle, to explore the relationship between temperature and transmission of coronavirus.

II. THE DATA

A. CoVCSD - Covid-19 Countries Statistical Dataset (Dataset A)

The datasets hold information about the cases and deaths from COVID-19 for multiple countries between January 22, 2020, to March 30, 2020, and there is a separate excel sheet for every country. The total size of this data is 8MB. This dataset includes 90 countries in total, which are Afghanistan, Albania, Algeria, Andorra, Angola, Antigua, Argentina, Austria, Azerbaijan, Bahamas, Bahrain, Bangladesh, Barbados, Belarus, Belgium, Belize, Spain, Denmark, Djibouti, Dominican-Republic, Ecuador, Egypt, Equatorial-Guinea, Estonia, Eswatini, Ethiopia, Finland, France, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Guatemala, Guinea, Guyana, Honduras, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Kazakhstan, Kenya, Kuwait, Kyrgyzstan, Laos, Latvia, Lebanon, Liechtenstein, Liberia, Libya, Lithuania, Luxembourg, Madagascar, Malaysia, Maldives, Mali, Malta, Mauritius, Mexico, Moldova, Mongolia, Montenegro, Morocco, Mozambique, Namibia, Nepal, Netherlands, New Zealand, Nigeria, Norway, Oman, South Korea, Tanzania, Thailand, Togo, Trinidad and Tobago, Tunisia, Turkey, Uganda, Ukraine, United Arab Emirates, USA. In excel sheet of each country with the following 31 attributes: date, country name, cumulative confirmed cases and cumulative deaths, daily cases reported and daily deaths, latitude and longitude for the country, average, minimum and maximum temperatures of that day, Wind speed, precipitation, fog, population, population density and median population, sex ratio, population above 65 years of age, hospital beds, available hospital beds/1000 people, confirmed COVID-19 cases/1000 people, number of males and females/1million people suffering from a lung disease, life expectancy, total COVID-19 tests conducted, outbound travels, inbound travels, domestic travels.

B. Novel Corona Virus 2019 Dataset (Dataset B)

This dataset has daily level information on the number of affected cases, deaths, and recovery from the 2019 novel coronavirus. The version data in this analysis is available from 22 January 2020 to 6 December 2020. The details of each column are as following: index number, Date of the observation in MM/DD/YYYY, Province or state of the observation, Country of observation, Time in UTC at which the row is updated for the given province or country, the cumulative number of confirmed cases till that date, the cumulative number of deaths till that date, the cumulative number of recovered cases till that date. The total size of this dataset is 11.85 MB

III. BACKGROUND

There are already many essays to analyze the correlation between temperature and the spread of coronavirus. In the case of China, the temperature has effects on death cases, confirmed cases, and suspected cases[1]. When the temperature between 2 and 20, the virus will spread faster. Besides, the areas in higher temperatures have a lower number of death cases.

IV. A HYPOTHESIS

With the knowledge that the temperature and climate may affect the spread of COVID-19, it follows that there could exist a relationship between temperature and COVID-19 spread.

The null hypothesis has been drawn:

The temperature does not affect the transmission of the virus.

V. DATA CLEANING AND PRE-PROCESSING

In Dataset A, the only cleaning thing that needs to do in data cleaning and pre-processing is to delete the missing values. Because the missing values are the names of countries or regions. If missing values are confirmed cases number, maybe can be replaced by the average of nearby value but the country and regions are missed, only can delete them. All the countries have their own data file, to simplify the data processing, all the files should be combined. In Dataset B, some names of provinces are missed which need to be deleted and the observation times of each country are according to their time zone, so the time needs to be standardized. Besides, the metrics of temperature are different in dataset B, converting Fahrenheit metric to Celsius metric.

VI. DATA PROCESSING

A. The Spread of COVID-19 Cases Across World

Using Novel Corona Virus 2019 Dataset to analyze the spread of COVID-19 across the world. However, all the data are sorted by observation time and specific regions of each country, first thing is calculating the cumulative case data of each country and each observation date. Because the case data are from 22 January 2020 to 6 December 2020, showing the case distribution every day is impossible, then only choose one day in each month to represent the distribution of that month and plot by the logarithm of cumulative confirmed cases. Figure1 shows the distribution of the confirmed cases all over the world from January to December. The plots in the first line from left to right are January, February, March, and April separately, and the plots in the second line from left to right are May, June, July, and August. The order of the rest plots follows the same rule. The darker color means the more confirmed cases. From Figure 1, in the first several months, China had the most cumulative confirmed cases and America had a few. In March, the coronavirus spread fast and it nearly reached every country in the world. Besides, Europe, America, and China faced a similar severe situation in March. From February to March, coronavirus only used one month to spread to the whole of Europe and South America. Then Europe and America as the new epicenter for the virus, where there is a rapid rise in COVID-19 cases in European Countries and The US, whereas the cumulative confirmed cases in China kept stable and the number of confirmed cases gradually spreads throughout the world. However, from the spread trend in the whole world all over the year can not find the effects of the temperature. Because the virus spread too fast with the spike in confirmed cases, even though the temperature does have some effects on the viral transmission, the initial numbers are huge and after using the log, the effects are hard to be found from the plots. Therefore, need deeper analysis.

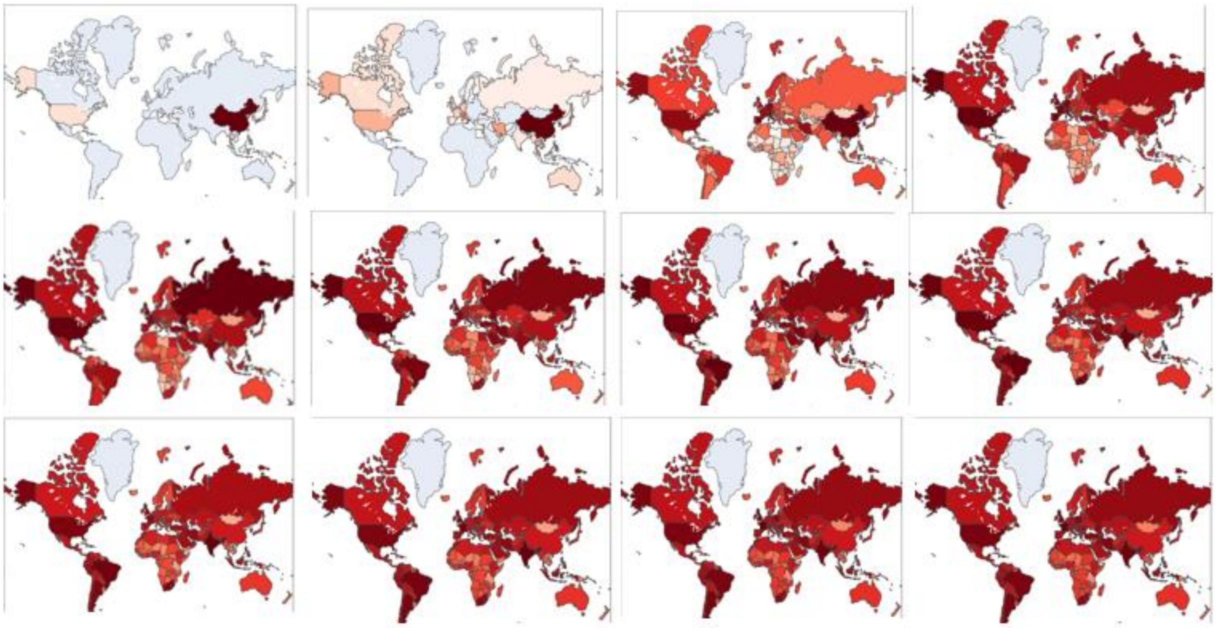


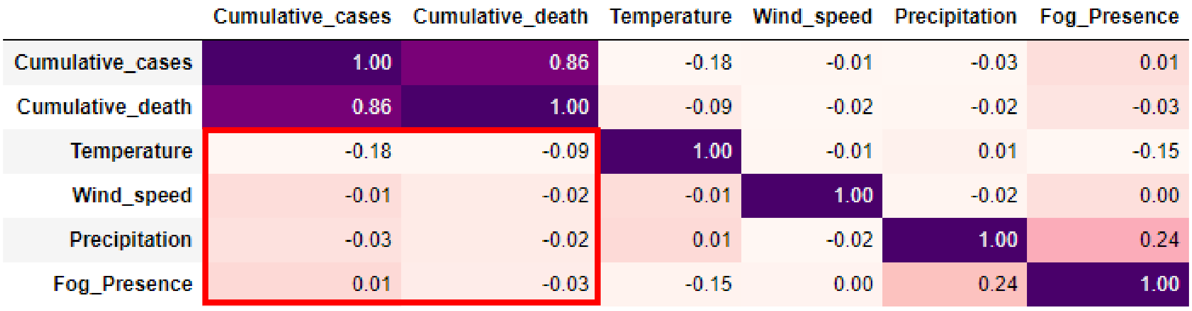
Figure 1. The cumulative confirmed cases in the whole world from Jan 2020 to Dec 2020

B. Explore Pearson product-moment correlation coefficient

In this step, to find the correlation between weather and virus spread, using CoVCSD - Covid-19 Countries Statistical dataset, because it includes many attributes about the weather. A Pearson product-moment correlation coefficient attempts to establish a line of best fit through a dataset of two variables by essentially laying out the expected values and the resulting Pearson's correlation coefficient indicates how far away the actual dataset is from the expected values. Depending on the sign of Pearson's correlation coefficient, which can end up with either a negative or positive correlation if there is any sort of relationship between the variables of a dataset. The formula for calculating Pearson's correlation coefficient as following:

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma X \sigma Y} \qquad (1)$$

In equation 1, X and Y are two random variables, σX and σY means standard deviation, $cov(X,Y)$ means covariance. In this dataset, there are four attributes related to weather and temperature: Temperature, Wind_speed, Precipitation, and Fog_Presence. And two attributes indicate the spread of the virus: Cumulative_cases and Cumulative_death. Therefore, calculating the Pearson correlation coefficient between four weather items and two virus items, the outcomes are shown in Figure 2.



	Cumulative_cases	Cumulative_death	Temperature	Wind_speed	Precipitation	Fog_Presence
Cumulative_cases	1.00	0.86	-0.18	-0.01	-0.03	0.01
Cumulative_death	0.86	1.00	-0.09	-0.02	-0.02	-0.03
Temperature	-0.18	-0.09	1.00	-0.01	0.01	-0.15
Wind_speed	-0.01	-0.02	-0.01	1.00	-0.02	0.00
Precipitation	-0.03	-0.02	0.01	-0.02	1.00	0.24
Fog_Presence	0.01	-0.03	-0.15	0.00	0.24	1.00

Figure 2. Pearson's correlation coefficient between weather attributes and spread attributes

The value of a correlation coefficient ranges between -1 and +1. The correlation coefficient is +1 in the case of a perfect increasing linear relationship, −1 in the case of a perfect decreasing linear relationship[2], and some value in the open interval (-1,1) in all other cases, indicating the degree of linear dependence between the variables. As it approaches zero there is less of a relationship. The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables. From figure 2, the absolute value of the coefficient of Wind_speed, Precipitation, and Fog_Presence is close to 0, indicating there is no relationship between them and the coefficients between temperature and Cumulative_cases and between temperature and Cumulative_death are -0.18 and -0.09 respectively. Even though the absolute values are still small, they show with the temperature rise, the confirmed cases tend to slow down. But the weak correlation only means there is no strong linear relationship. There are deeper analyses of correlation needed.

C. Find the relationship between temperature and confirmed case

From process B, it is clear that the relationship between temperature and confirmed cases is not linear. To show the correlation, choosing temperature as the x-axis, and confirmed cases as the y axis. Figure 3 shows all the coordinates clearly. When the temperature from -30 to -5, the confirmed case stays at a low and stable level. However, confirmed cases have a significant growth at the temperature from -5 to 17. The confirmed case is the highest at about 15. If the temperature keeps rising until 37, the number of a confirmed case decreases rapidly and stay stable. From this plot, the temperature does affect the spread of the virus. This can also explain why in process A, we cannot find any case increased or case decreased trade relate to temperature. In Figure 1, China, Europe, and America are three epicenters for the virus. In China, the confirmed case number nearly not changed after March, because of the efficient lockdown police, even though the temperature was suitable for spreading, there are no increased cases in China. However, in Europe and America, February and March were the breaking times, just like China in January. Though the temperature may be one of the reasons for rising rapidly, the main factor is the government and citizens have not started to take measures to fight the epidemic. After March, the temperature led to people easier to contract the virus, nevertheless, the huge initial confirmed cases number and the plots use the logarithm, the effect of temperature is hard to be found.

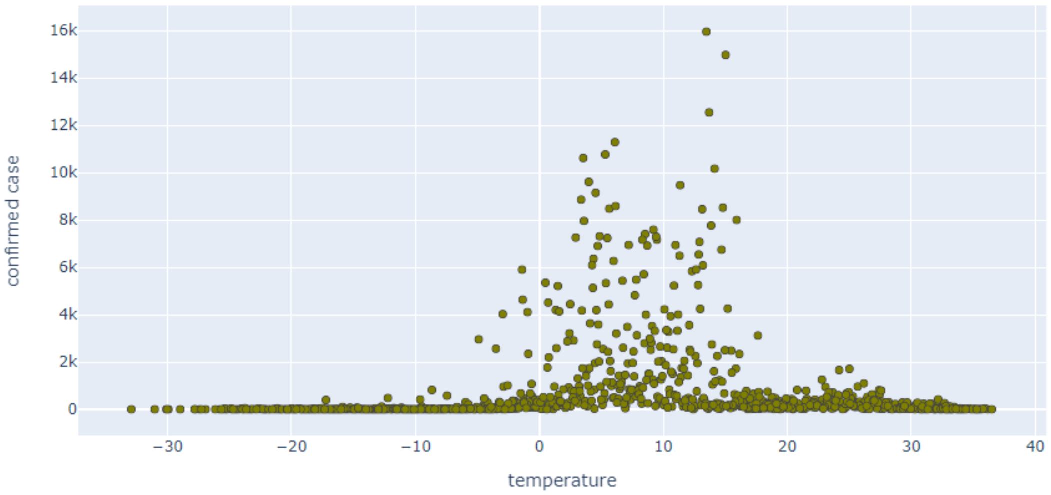


Figure 3 Scatter diagram of temperature and confirmed case

D. hypothesis testing

Even though from the processes above, we can find some relationship between temperature and spread of the virus, it is necessary to do hypothesis testing. To make sure this difference is statistically significant. Randomly choosing a sample group from the temperature from -5 to 17, with sample size is 250, and also choosing a test group from the temperature the rest of ranges with sample size is 250. Hence if P-value > 0.05 we can safely accept our null hypothesis that there is no statistically significant, the higher confirmed cases at the temperature from -5 to 17 can not indicate the temperature can affect the spread of the virus. The p-value of the t-test is 0.285, Since we get p value > 0.05 we can accept the null hypothesis and can conclude, no statistical difference is present between the two datasets and the sole effect of temperature on the spread of COVID-19 can be safely rejected. However, the idea of the spread of COVID-19 across a certain range of temperature needs more dataset and statistical testing to come up with a substantial conclusion.

VII ANALYSIS

A. Process analysis

The whole analysis aims to figure out will temperature influence the spread of the coronavirus. At first, hoping to do a general analysis of the spread situation in each country (in VI-A, process A), but because of the huge initial number and plot in logarithm, it is hard to find the temperature effects. Then in process B, choosing Pearson’s correlation coefficient to do some deeper analyses. Calculating four Pearson’s correlation coefficients of temperature, Wind_speed, precipitation, and Fog_Presence. The coefficient of temperature is -0.18, which means there is a negative relationship between them. But the absolute value of -0.18 is too close to 0, indicating the linear relationship between them is weak. And the other coefficients show there are no effective of Wind_speed, precipitation, and Fog_Presence.

Two possible reasons lead to the weak coefficient in process B, first, there is no relationship between temperature and coronavirus. Second, the relationship exists, but not linear. Then the process C is needed, the plot with temperature on the x-axis and confirmed cases on the y axis indicating that when temperature between -5 to 17, the confirmed case increased a lot. It seems the reason should be the second one, there is no linear relationship between temperature and virus spread, at a certain range of temperature, the virus will be easier to spread. Cannot give the conclusion only from this plot, the t-test is needed to test the difference of confirmed cases between the temperature range of -5 to 17 and others temperature is statistically significant. In process D, randomly choose 250 samples from the temperature from -5 to 17 and others respectively to do the t-test. The p-value of this t-test is 0.28, p-value > 0.05, the null hypothesis can be accepted. There is no significant difference between the two groups, indicating the relationship found in process C means nothing. The conclusion is the temperature does not affect the spread of coronavirus.

B. Conclusion analysis

The final analysis was beyond my expectation, temperature does not affect the spread of coronavirus, which differs from the known truth theory[1]. the temperature does not affect the spread of coronavirus. The whole analysis process is logical and meticulous, so why this thing happened? After reading two references again, I find the reasons. In the first reference, all the results it gives only from the scatter plots. It did not do any t-test to true the differences between temperatures are significant. Besides, there are some interference factors it did not notice. For example, Inter Mongolian has low temperature and low confirmed cases. This not only because of the temperature, but Inner Mongolia is also a place with a very low population density, most people there live in prairie, which means People live far apart from each other. And it hard for the virus to spread. Moreover, after the outbreak of the virus in Wuhan, the government decreed that no one could enter or leave Inner Mongolia to provide the virus spread. Maybe it seems not to be professional with using statements not data, but I lived in Inner Mongolia during the whole coronavirus period, and I do know this place. The regional characteristics and government controls of Inner Mongolia help a lot in controlling the virus. Therefore, in the first reference, many interference factors like regional characteristics and government restrictions are overlooked, and the only mind the plot about temperature and the confirmed case is not convinced. Even though the conclusions of the first reference are not convinced, still find a lot of references to prove the temperature of virus spread. Comparing my analysis with theirs, the main difference is my data is not enough. My confirmed cases and temperature data are only last until March. If using the data of the whole year, maybe the result will differ.

VIII CONCLUSIONS

Even though there are some differences shown on the plot in process C, the t-test result indicating it is not statistically significant indifference. The null hypothesis is accepted, the temperature does not affect the spread of coronavirus.

A. Limitations

There are three limitations to my analysis. Firstly, my analysis also does not consider the interference factors, like the first reference. Regional population density, government controls, COVID-19 detection capability, population mobility, and treatment ability are all can affect the confirmed cases and death cases. However, these factors are hard to be quantified, which increases the difficulty of analysis. Besides, the limitation of the size of the dataset also influences the outcome. Only using the data from January to March, can not show the temperature variation. Finally, to figure out the effects of temperature, humidity, and wind speed can not be ignored. The spread of the virus is related to the wind speed and the density of coronavirus in the air, which is the mental factor of people infected with coronavirus, which is also affected by wind speed. Therefore, these weather factors also should be concerned.

B. Further steps

According to the analysis, three ways can make it more complete. The first one is also the easiest, finding a larger dataset to do the analysis again. The second way, using the data from the same region, which can make sure the interference factors hard to be quantified are same, even though they still exist, they are same then the effects from them can be ignored. It is worth noting that the region should not be too large, If the region spans a large latitude and longitude, then the effects from interference factors will differ. And the last one, combining biological analysis and geographical analysis. After analysis of the activity and infectivity of coronavirus in different temperature in the biological lab, which can help avoid all the interferences from other factors, then combine the biological analysis outcomes with the geographical analysis together, to find out the temperature effects on transmission of the virus and get a convincing result about temperature and confirmed cases.

Reference

[1] Siddiqui, Mohammad Khubeb & Morales-Menendez, Ruben & Gupta, Pradeep & Iqbal, Hafiz & Hussain, Fida & Khatoon, Khudeja & Ahmad, Sultan. (2020). Correlation Between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases based on Machine Learning Analysis. Journal of Pure and Applied Microbiology. 14. 10.22207/JPAM.14.SPL1.40.

[2] Dowdy, S. and Wearden, S. (1983). "Statistics for Research", Wiley. ISBN 0-471-08602-9 pp 230

[3] https://www.kaggle.com/aestheteaman01/covcsd-covid19-countries-statistical-dataset

[4] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset