Investigation on the relationship between Covid-19 spread speed & population density in China and America

Geographic Information System and Science
Centre of Advanced Spatial Analysis

GitHub repository URL: https://github.com/TTonsss/GIS_Final.git

Word count: 2942

2021-1-7

1. Introduction

In 2020, the whole world has suffered from the global pandemic, Coronavirus Disease 2019 (Covid-19). The first Covid-19 case was found in Wuhan, China, in December 2019, then China became the first country that needs to deal with this virus. Covid-19 has since spread worldwide through 2020, and became an ongoing global pandemic. The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes (WHO 2020). So, it makes the virus really easy to spread.

With the significant spread speed, the confirm cases has raised quickly and this of course increase the death cases which is the result that everyone does not like at all. This report is investigating the dataset of Chinese Provinces confirm cases from January to October in 2020 and American States confirm cases from January to December in 2020. And the purpose is to find out if there is any the relation between the Covid-19 spread speed and population density in these two different countries separately.

2. Literature Review

A lot of analysts put attention on the topic of the disease spread. There are many factors that affects the transmission and fatality of covid-19 in metropolises such as the wind speed, the longitude – similar with the temperature, GDP per capita, etc (W.Cao 2020). The total number of Covid-19 death report by European country up to 31 July 2020 demonstrates the peak of death in countries with 1.5 million population is about 2500 per week. In countries with 1.0 million population, it is just above 1500 per week and in countries with 0.5 million population, the death per week is about 750 per week. Therefore, the population is positive correlated with the total death cases.

2.1 Why choose population density as the factor?

The density population also demonstrates positive correlation to the total confirmed cases and mortality in India (Arunava Bhadra, 2020), and Algeria (Nadjat Kadi, 2020). A research in Turkey gives out the result that the spread of Covid-19 is not only affected by the population density but also considered the speed of wind catalyses the epidemics (Hamit Coskun, 2020).

Looking through all of those factors, since the research is based on the metropolises, the cities with high population density are more concerned by experts. So, take a step back, the population density is more likely to relate to the confirm cases and death cases. In the same country, because of the geometry, the affect factors do not have much differences. But the population density of different cities is significantly difference from each other which is caused by the urban diversity and the different economic system. It is clearly more efficient to investigate the relationship between population and Covid-19 base on the researching of one specific country.

2.2 Why choose transmission speed?

From all of the research above, the number of total cases is used as the dependent variable. The amount generates the facts that Covid-19 has brought to us. It is a good choice to show what result Covid-19 has caused. It will let the government make decisions on sending medical personnel, allocate funding and prepare plenty room for the patients.

However, in this case, a factor which bases on time is needed to generate the influence of the virus rather than a factor of amount. Thus, the spread speed is considered. The value of how fast the epidemics spread is able to directly illustrate the infectious of it in the specific country. This may be caused by the different culture, different environment or different understanding of preventing disease. The significant transmission speed is much clearer to show the trend of the coronavirus spreading and it is easier to draw attention to the government to prevent the disease and make faster decisions on the area which is need to be quarantined. Therefore, investigating and analysing the data of Covid-19 spread speed is more helpful on the forecast and inhibit the catalysis of the pandemic.

2.3 Why choose America and China?

China, as the first country suffered from the coronavirus disease and a developing country with diversity of population distribution in the whole country, it is worth to investigate the data of Chinese Covid-19 situation. Since it is assaulted by the epidemic without any alert, the disease spread in a more natural way at the first half period. On the other hand, America is a developed country with full alert and well prepared for the pandemic. The population density in different states is also well distributed (Darryl Cohen, 2015). And the first case in America is discovered on 21st January 2020, 2 months later than the first case in China. It will be interesting to investigate the relationship between the spread speed of Covid-19 and the population density in these two different background huge countries, but both of them contain various levels of city with different population density which provide enough diversity of the dataset.

2.4 Why only investigate confirmed cases?

The virus lethality rate is a characteristic of a kind of virus and it is not going to change rapidly in a country. So, the death cases are strong positive correlated to the confirmed cases and it is less concerned under this topic, since the spread speed of Covid-19 is reflected more intuitively by the confirmed case, as it means the end of a virus transmission.

3. Methodology

3.1 Chosen area when calculating population density

In China, the way of calculating population density is by provinces. And in America the chosen area is state. The investigation on the data in this topic is based on the whole country. If choose metropolitans or counties as the area to gather data, it is not able to generate a good map with cities or counties in the country map. If to do so, the area of each unit will be too small and there will be too many elements in the map. That will lead to an unclear declaration of information for audience from the map. Also, it is needed to mention that the final purpose is to let governments make decisions easier to control the pandemic nationwide. So that provinces in China and states in the USA are more related to the whole country and lead the result of this analysis makes more sense.

3.2 Define and Convert population density

The population density is known as a measurement of population in the unit area. The units of population density in China and America are different. In China the density is measured in inhabitants per kilometre square. But in the USA, it is measured in inhabitants per mile square. To prepare for the later comparison, it is better to convert them to the same unit. So, in this case inhabitants per kilometre square is chosen to be the standard unit and the value of American population density is divided by 2.5900 to convert to the standard, shown as below:

$$\frac{inhabitants}{mile^2} = \frac{inhabitants}{(1.60934 \times km)^2} = \frac{1}{2.5900} \times \frac{inhabitants}{km^2}.$$

3.3 Calculate virus spread speed – Basic

In order to achieve the purpose that demonstrating the spread speed clearly, the average increase ratio is used in this case. The average increase ratio is defined as the cumulative confirm cases increase in each time step divided the effective period. Effective period is defined as the amount time that confirmed Covid-19 case exits in the area.

$$\overline{Pd_i} = \frac{\sum_{j=1}^{t-1} \frac{C_{i,j+1} - C_{i,j}}{C_{i,j}} * a_{i,j}}{t - 0^{a_{i,j}}}, i \in n, a_{i,j} = \begin{cases} 1 & \text{if } C_{i,j} > 0 \\ 0 & \text{if } C_{i,j} = 0 \end{cases}.$$

Where, n: Number of provinces in China or states in America. t: total time period that contains all data. Pd: Population density in area i. C: the matrix of total confirmed cases up to date.

$$C_{n,t} = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,t} \\ a_{2,1} & a_{2,2} & \cdots & c_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n,1} & c_{n,2} & \cdots & c_{n,t} \end{pmatrix}$$

$a_(i,j)$: the binary variable depends on the value of elements in matrix C. Adding the binary variable is helping the calculation to ignore the situation with 0 confirmed case which is not needed to count into the efficient time period. Therefore, if this situation happens, the rate of increase should times 0 and the total time period should minus 1 at the same time.

3.4 Calculate virus spread speed – With time step

During the peak of the epidemic, the value of confirmed cases changes dramatically and so does the rate of increase cases. This might cause by the recovery of patients, the immediate quarantine of governments or the big events that have been hold locally. This will bias the data and the result will demonstrate less accurate on the embodiment of the local Covid-19 spread speed. Thus, the original time step which is one-day time step is not fit for this analysis. How about one-month time step? One month might be too long for observing the cases, since the incubation period of Covid-19 is in 2 weeks and the recovery time is different depends on how sick is the patient at the first place. In some area the virus may vanish in 1 months so that nothing can be observe if the time step is that long. To discover whether different time step influence the result or not, there is an alternative way to calculate the spread ratio of increase cases include changeable time step:

$$\overline{Pd_i} = \frac{\sum_{j=1}^{t-s} \frac{C_{i,j+s} - C_{i,j}}{C_{i,j}} * a_{i,j}}{t - 0^{a_{i,j}} * s}, i \in N, a_{i,j} = \begin{cases} 1 & \text{if } C_{i,j} > 0 \\ 0 & \text{if } C_{i,j} = 0 \end{cases}$$

Where s is stand for the value of the time step, and other variables stay same as the basic method.

3.5 Linear regression

Linear regression is a linear approach to modelling the relationship between a dependent variable and one or more independent variables (Arunava Bhadra, 2020). The independent variable in this case is the population density in each area of China and America, and the dependent variable is the rate of increase in confirm cases. The model generates out a line which is best fitted (minimum sum of differences from data point to this line) with the function

$$y = ax + b,$$

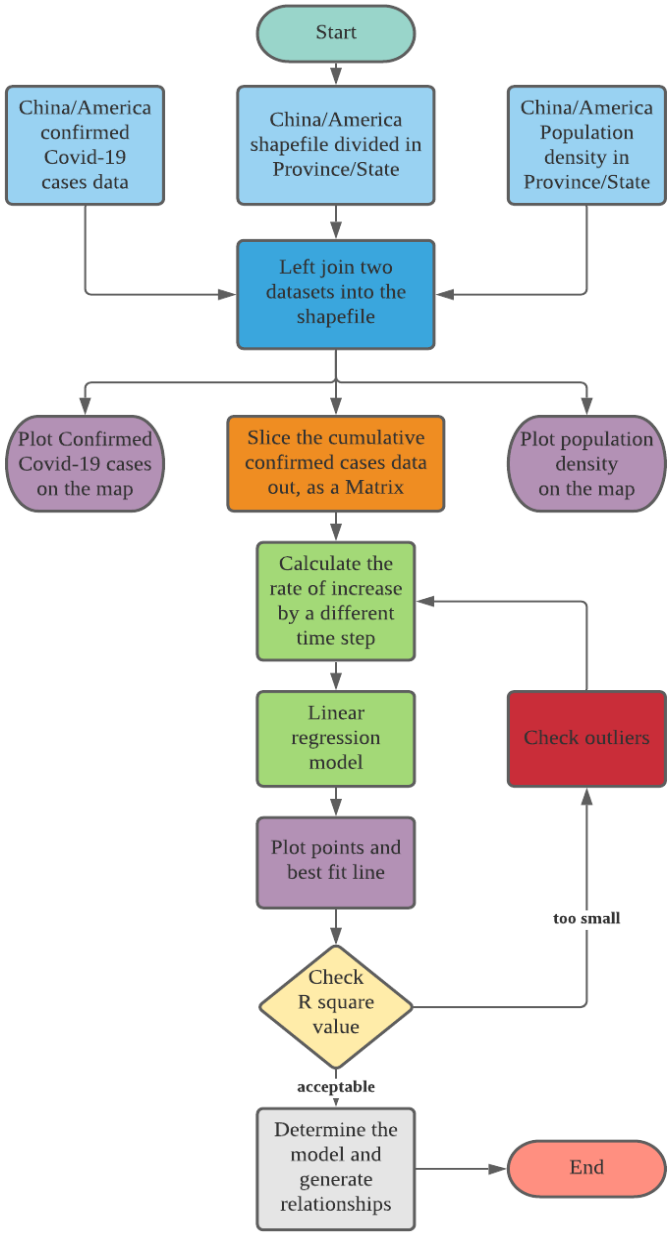
where y is the dependent variable, x is the independent variable, a is the slope, and b is the interception that the line across on y-axis. In the present work, y is the increase ratio, x is the population density. a is the acceleration of the confirmed cases under different population density.

The coefficient of determination (R^2) quantifies the amount of variability in dependent variable explained by the model and is defined by the relation

$$R^2 = \frac{SST}{SSE} - 1,$$

where SSE is the sum of squared errors (squared residuals) $SSE= {i=1}^{n}(y_i-y_ie)^2 $, *and SST is the sum of squared variation in dependent variable about its mean* $SST={i=1}^n(y_i-{y_i })^2 $. With the increase in number of regressions, the value of R^2 always increases. For a meaningful measure of goodness of model fit, adjusted R^2 is used instead of considering the number of explanatory variables in the model (Arunava Bhadra, 2020).

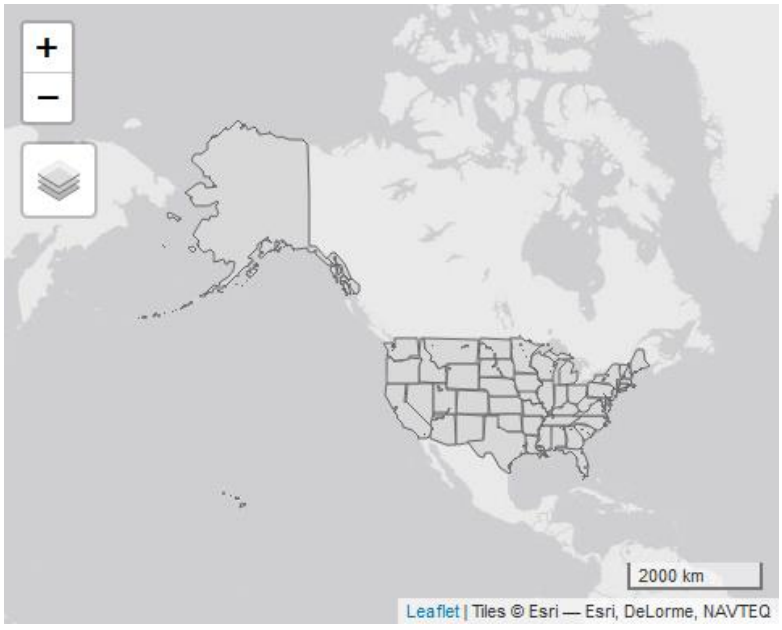
3.5 Flowchart



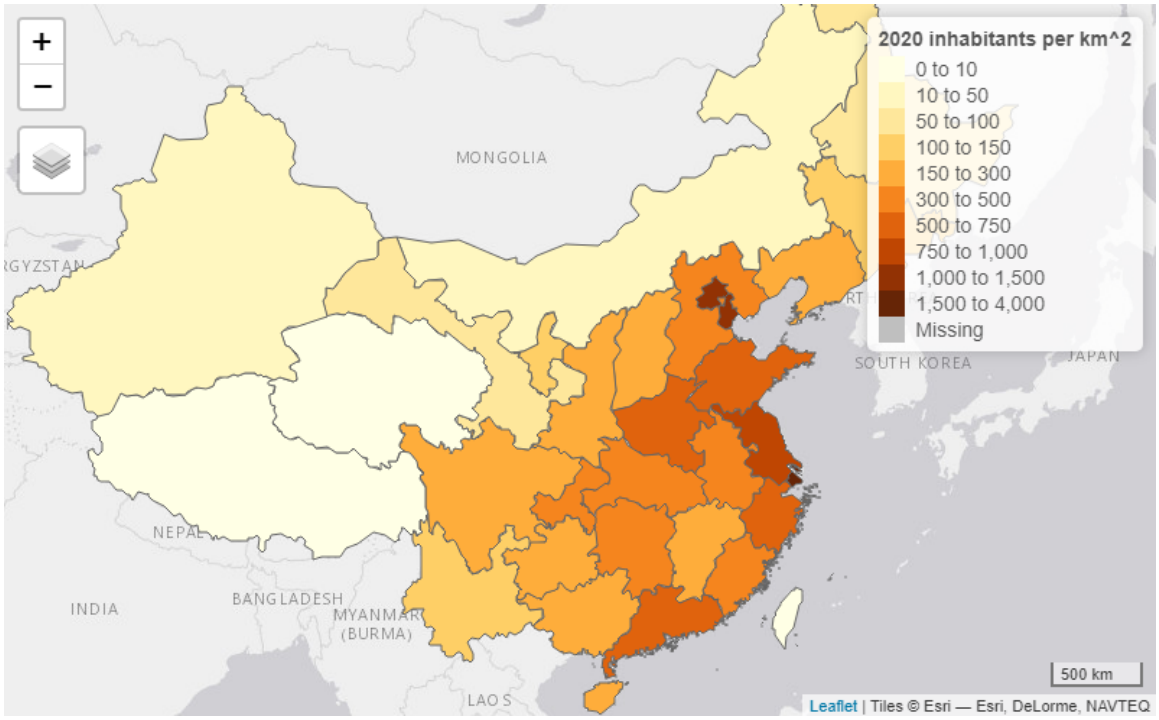
4. Result

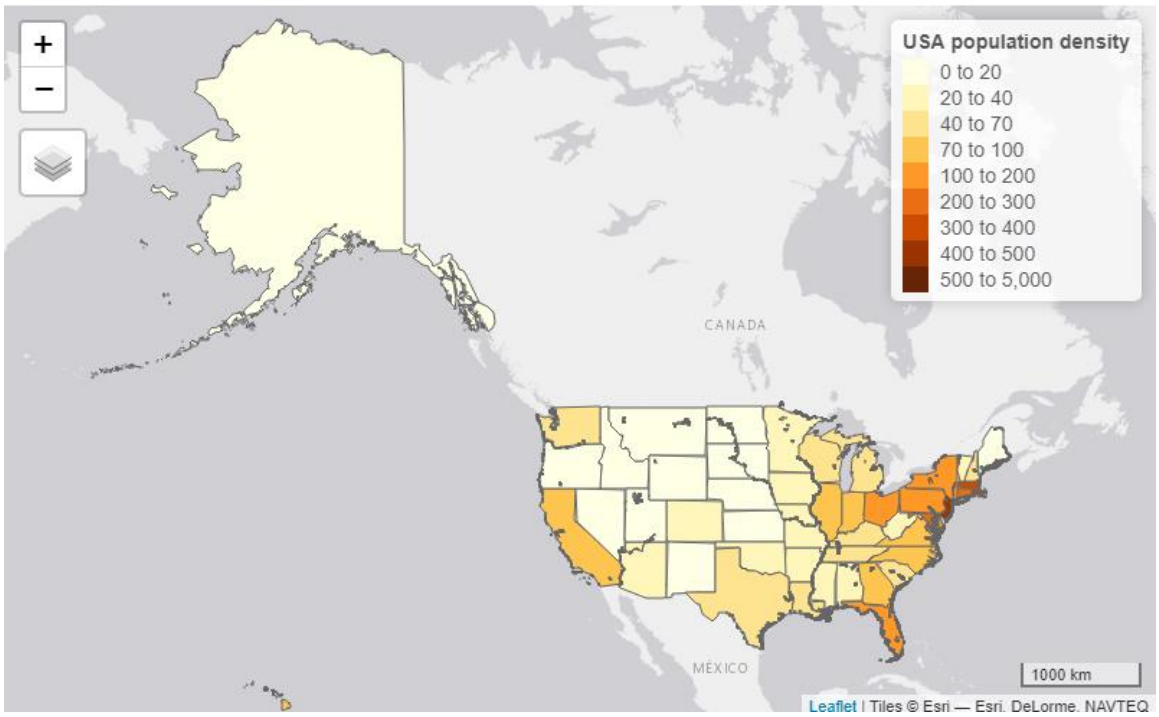
4.1 Prepare

Firstly, Check the shape files are working, and the base maps are generated below.

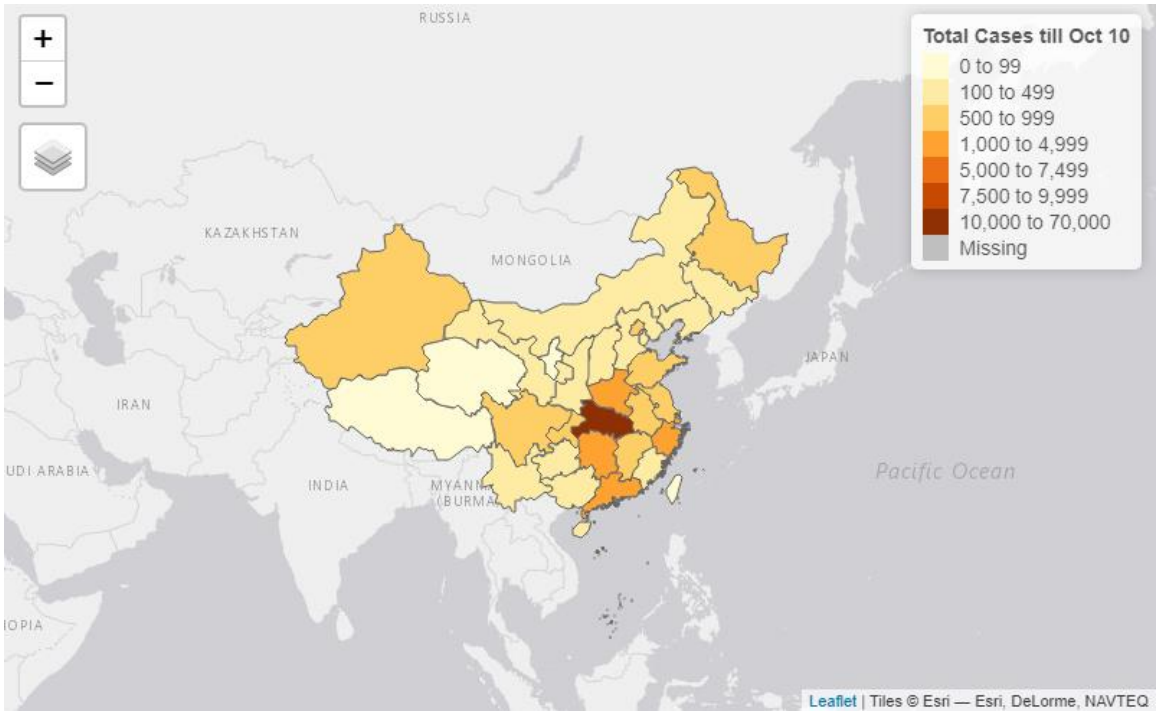


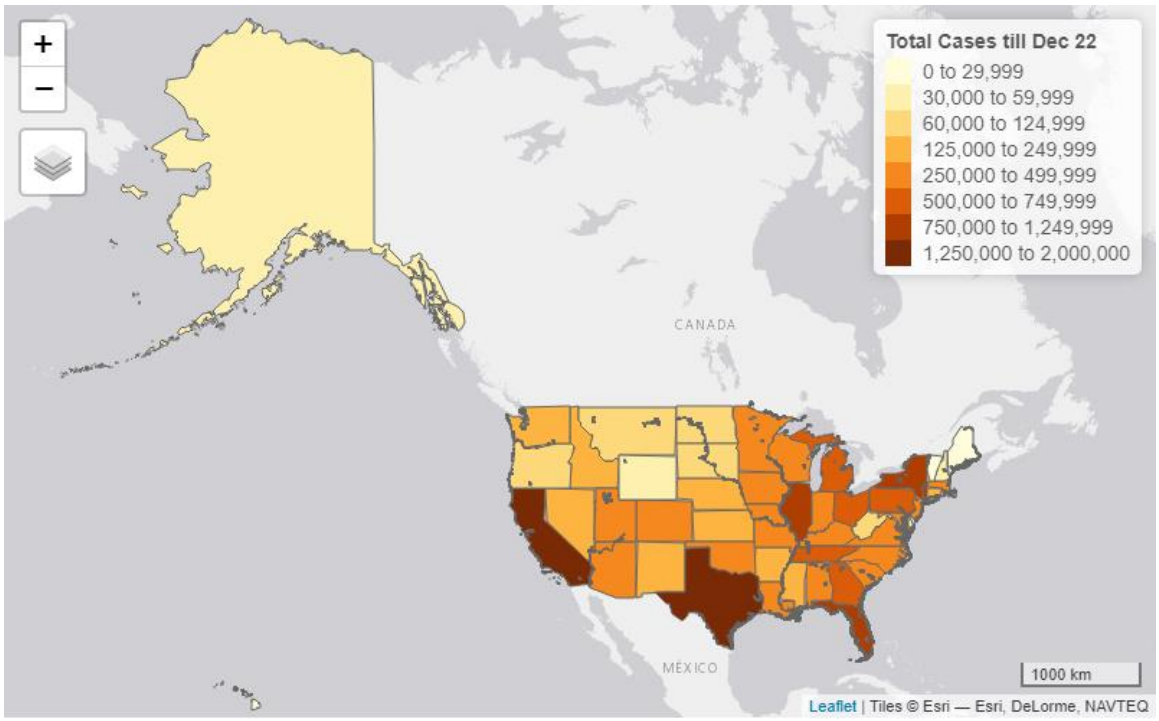
Then After left join the shape file with the population density in each province or state. And the population density in 2020 in China and the USA are shown on the map below.





What comes next is loading the data of confirmed Covid-19 cases in to R and left join the data into the shape file. Then it is able to plot the total number of confirmed cases on the map, in this case the statistic ends on 10th October in China and 22nd December in America.



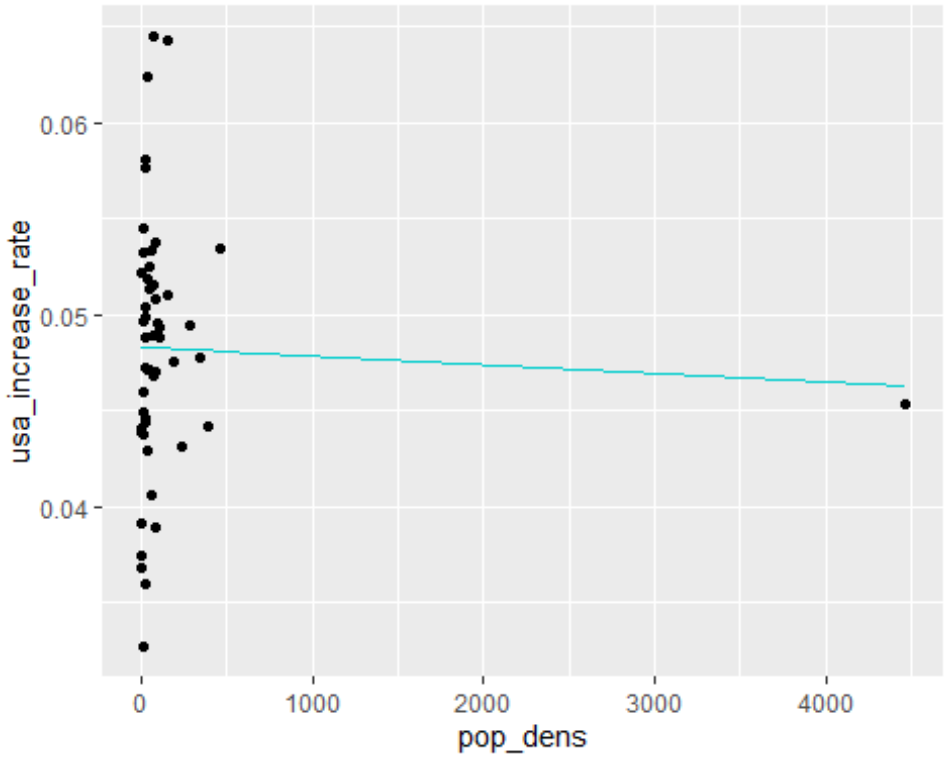


Then, calculation of the average rate of increase in confirmed cases is needed.

4.2 Case: American

America is selected as an example for finding the appropriate time step. The time step of 1 day the rate of transmission is calculated. After this calculation, use the rate and the population density to make the linear regression model.

```
Call:
lm(formula = usa_increase_rate ~ pop_dens, data = .)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0155724 -0.0039271  0.0005248  0.0034118  0.0162002

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.832e-02  9.697e-04  49.831   <2e-16 ***
pop_dens    -4.437e-07  1.519e-06  -0.292    0.771
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.006692 on 49 degrees of freedom
Multiple R-squared:  0.001738,  Adjusted R-squared:  -0.01863
F-statistic: 0.0853 on 1 and 49 DF,  p-value: 0.7715
```



As we can see the $R^2 = 0.001738$ which is far too small to say this is a good model. And we also detected an outlier which is District of Columbia contains the population density about 4467 inhabitants per km^2. It is removed in the next test of time step. So a new time step is needed, and this time try to scratch the data every months.

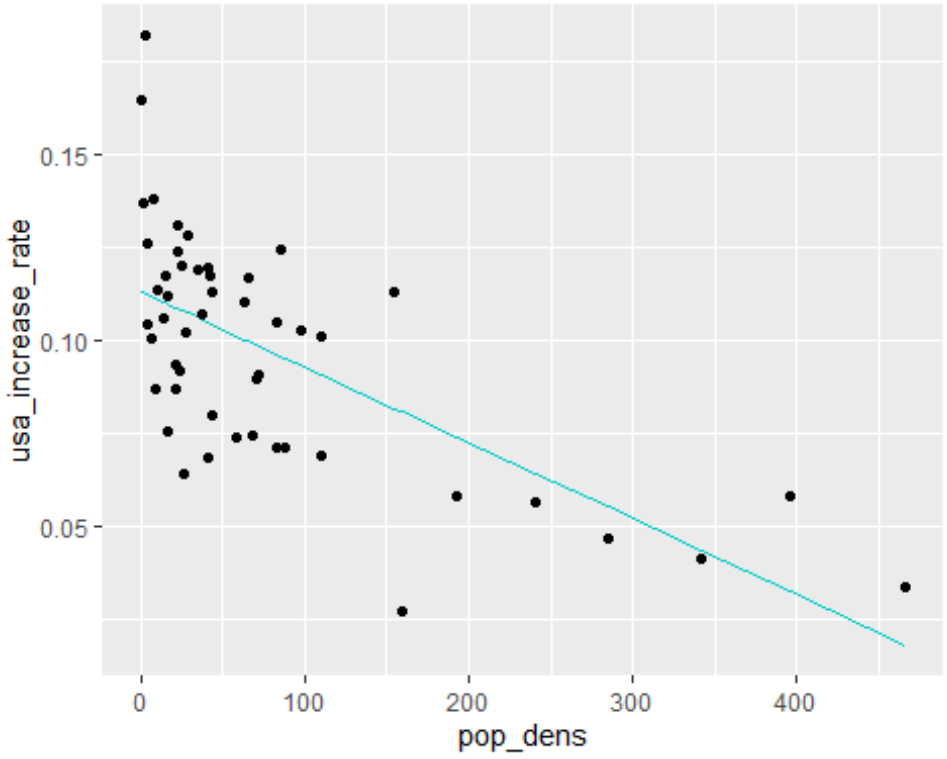
```
Call:
lm(formula = usa_increase_rate ~ pop_dens, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-1827.2  -181.4     9.2    96.2  7021.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -167.347    196.777  -0.850   0.3993
pop_dens       4.074      1.533   2.658   0.0107 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1103 on 48 degrees of freedom
Multiple R-squared:  0.1283,    Adjusted R-squared:  0.1101
F-statistic: 7.063 on 1 and 48 DF,  p-value: 0.01066
```

This time $R^2 = 0.1283$ which is better than calculate daily average rate of increase, but the value is still too small and make the regression a bad model. Now, set the time step as 1 week and do the whole process again.

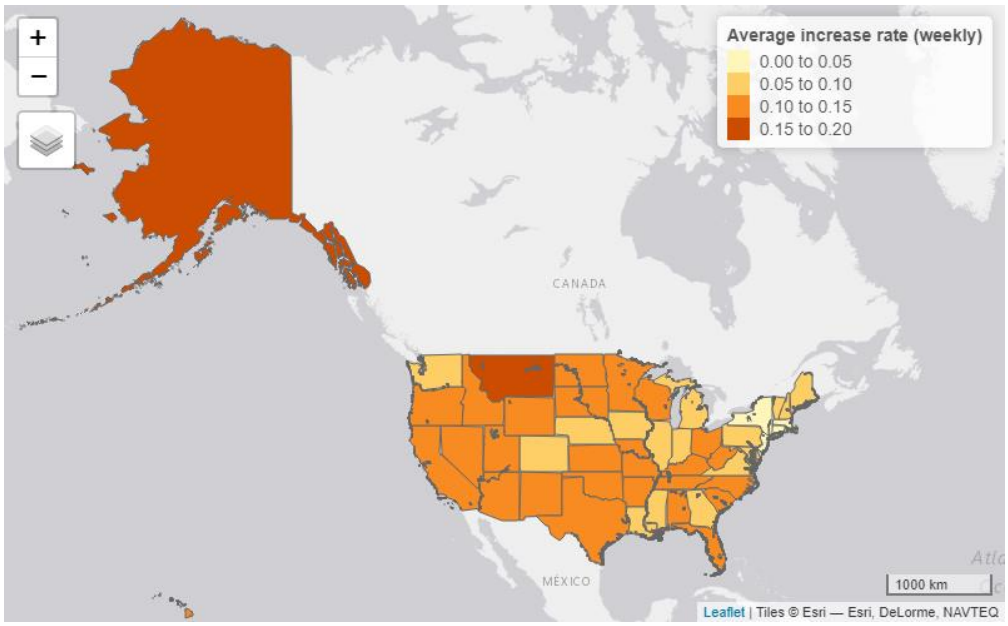
```
Call:
lm(formula = usa_increase_rate ~ pop_dens, data = .)

Residuals:
   Min     1Q Median     3Q    Max
-63.33 -15.69  -9.56  -3.37 317.46

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.32854   10.07892   1.025    0.311
pop_dens     0.01956    0.07852   0.249    0.804

Residual standard error: 56.5 on 48 degrees of freedom
Multiple R-squared:  0.001291,  Adjusted R-squared:  -0.01952
F-statistic: 0.06207 on 1 and 48 DF,  p-value: 0.8043
```

The $R^2 = 0.001291$ which is still small, so the chosen period is considered as a main factor that cause this happens. Then changing the total observation period is necessary since after the peak of the transmission, the rate will reduce quickly but it is still greater than 0, so ignore this useless period will make the small values influence less on the data.

```
Call:
lm(formula = usa_increase_rate ~ pop_dens, data = .)

Residuals:
      Min        1Q    Median        3Q       Max
-0.053516 -0.016095  0.001855  0.014390  0.069416

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.134e-01  4.276e-03  26.512  < 2e-16 ***
pop_dens    -2.044e-04  3.331e-05  -6.137 1.55e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02397 on 48 degrees of freedom
Multiple R-squared:  0.4396,    Adjusted R-squared:  0.428
F-statistic: 37.66 on 1 and 48 DF,  p-value: 1.549e-07
```



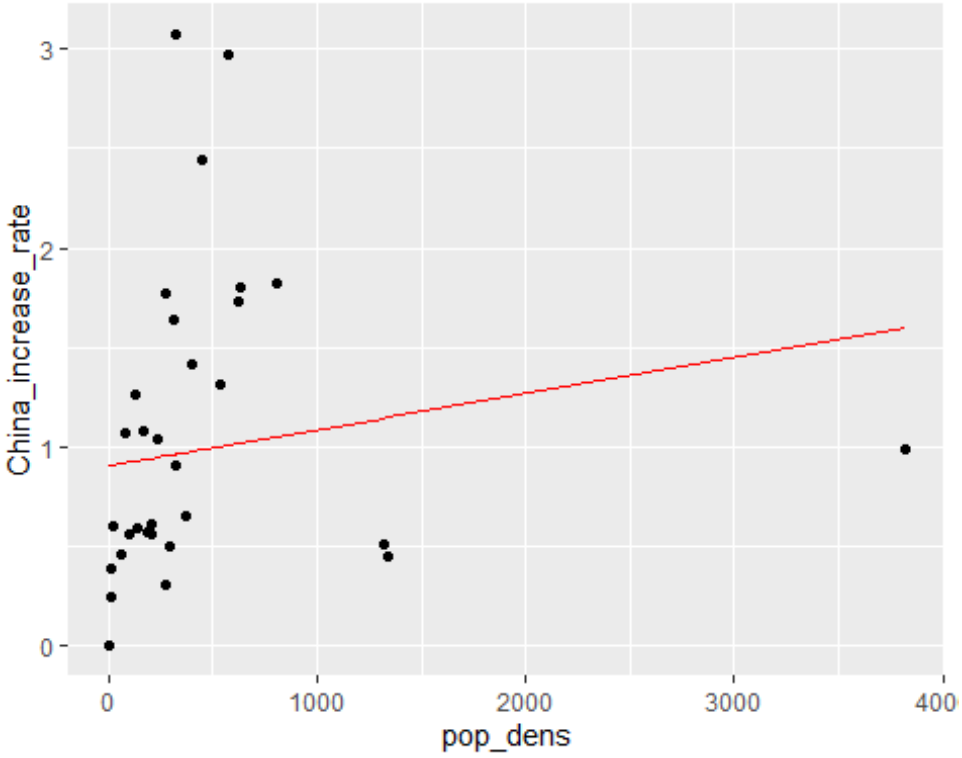
After selecting the months around the peak of the epidemic, the time period is from May to December. And the summary of the linear regression model gives out $R^2 = 0.4396$ which is acceptable. So this model can be used to describe the relationship between the average weekly increasing rate and the population density in each state of America. The slop of the best fit line is $-2.044 * 10^{-4}$ and the interception on $y - axis$ is 0.1134. For the better spatial visualisation, the increase is plotted on the map.



4.3 Case: China

Now, for the case in China, the one-week time step has been determined. Since the pandemic happens in China at the very beginning, so the time period is selected to be from January to June. By doing the same process as the American example, the detail of the Linear regression is listed below.

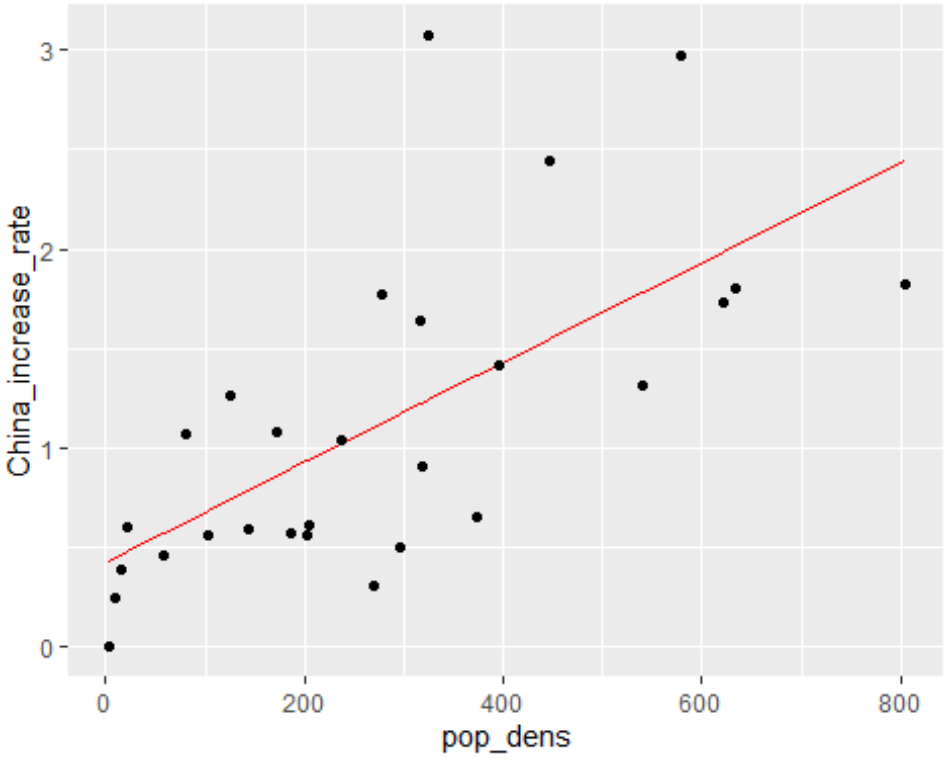
```
Call:
lm(formula = china_increase_rate ~ pop_dens, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9030 -0.5874 -0.3245  0.4152  2.1139

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.9024827  0.1621747   5.565 3.84e-06 ***
pop_dens    0.0001827  0.0002042   0.895    0.378
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8037 on 32 degrees of freedom
Multiple R-squared:  0.02441,   Adjusted R-squared:  -0.006073
F-statistic: 0.8008 on 1 and 32 DF,  p-value: 0.3775
```



The value of R^2 is 0.02441 which leads to a failed model. By looking at the graph, it is easy to discover there are 3 outliers in the data which are Beijing, Tianjin and Shanghai, these 3 areas are actually cities but with the same administrative level as

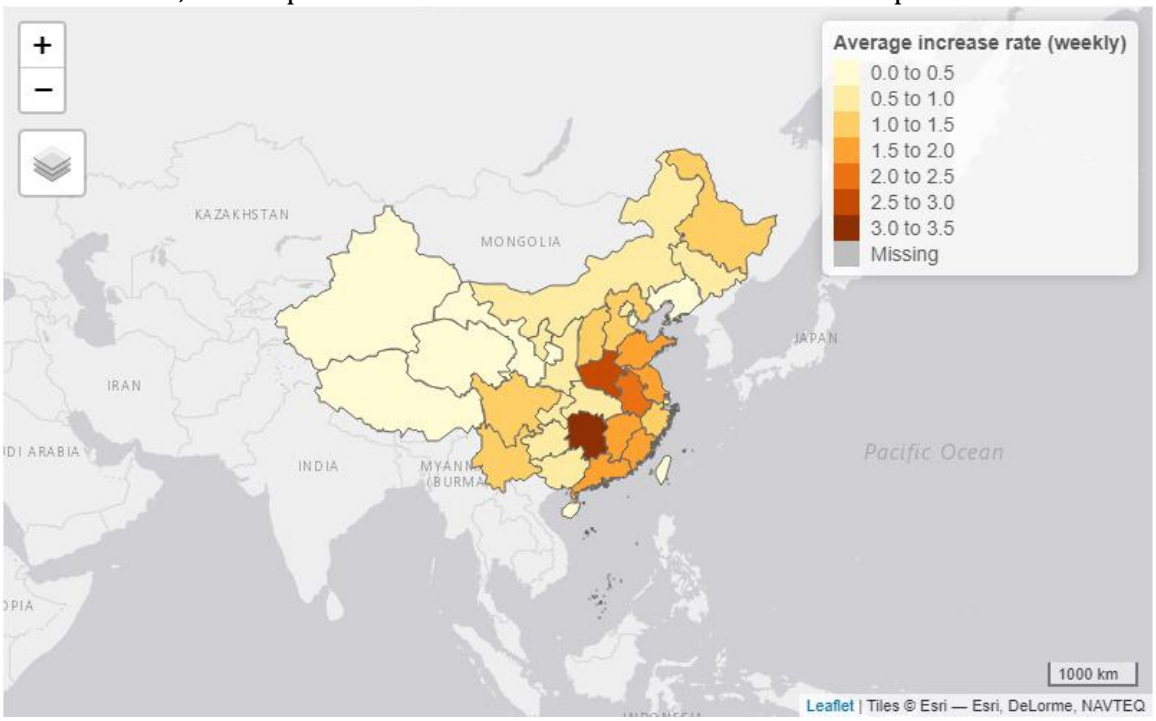
province and also the population density is extremely high. After removing the outliers, the new linear regression is like this:

```
Call:
lm(formula = china_increase_rate ~ pop_dens, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8033 -0.3389 -0.1625  0.2753  1.8382

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.423491   0.191160   2.215 0.035699 *
pop_dens    0.002514   0.000552   4.553 0.000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6076 on 26 degrees of freedom
Multiple R-squared:  0.4437,    Adjusted R-squared:  0.4223
F-statistic: 20.73 on 1 and 26 DF,  p-value: 0.0001092
```



Here $R^2 = 0.4437$, with the first province of the Covid-19, Wuhan, which may bias the data, the value of R^2 is acceptable. The linear model has the slope of 0.002514 and the $y - interception$ of 0.423491. After left-joining the shapefile with the

increase rate, the map is able to demonstrate the rate in all of the provinces in China.



5. Discussion

5.1 Relationship between total cases and population (density)

Looking at both of the China and America, the sum of confirmed cases at the end of the period is highly correlated to the population density, except few specific provinces or states which has the right measures to deal with the epidemic. In the area with less manual intervention, Covid-19 spreads naturally to a significant percentage of the population in that area. The area contains more population or more population density becomes disaster area more easily without quarantine at the end of the period. A high number of total confirmed cases with a stable lethality rate of Covid-19 leads to a big amount of death cases.

5.2 Relationship between spread speed and population density

It is easy to let people think that the spread speed is definitely connecting closely to the population density. The Chinese case illustrates this directly. The spread rate of the virus increases about 50% when the population density raises 200 inhabitants per km^2 which shows that the population density is a major factor of Covid-19 transmission. However as we known, Wuhan is the first quarantine province with the largest amount of cases in China, the rate of increased cases is 0.96 which is lower than the surrounding provinces. This is happening because the government quarantine the area immediately and force residents to stay at home which was efficient respective to the pandemic situation. But in the rest province around

Wuhan, Covid-19 transmitted in the “natural” way as the discuss above, the most rapid growth on the total cases happens in Hunan where the rate is above 3.

But this is the case that the whole world is focusing on. So, talking worldwide, the human intervention will also raise while the virus is spreading. The increase rate could reduce by it. That would cause by the local authority, the government or by attitude and behavior of all of the residents. The situation in America clearly shows that states with the high population density inhibit the growth of the confirmed number better than the states with low population density. The reason of it is America is alerted by the pandemic in China, the knowledge of preventing virus is also known well by the residents and the average population density in American states is less than China.

5.3 Development

The spread speed of Covid-19 is only calculated by the confirmed cases. Since the dead and recovered people are not participating the transmission, considering the death and recovered cases while doing calculation will improve the result and the regression model to make them more accurate.

6. Conclusion

Overall, American population density has negative correlation with the virus spread speed because of the well preparation and the relatively low population density compare to China which makes residents easy to prevent it. And Chinese population density has positive correlation with the virus growth speed in provinces. Even China with a rapid transmission, with a strict control by government, the total case has been restricted at a low level. The coronavirus-2019 is still damaging the world, and some cases of Mutant virus are found. It is possible that in the future, the vaccine will stop working for the mutant virus and countries all over the world will need start preventing it. It is necessary to keep focusing on the spread speed and discover more factors which can affect it.

Reference:

Nadjat Kadi and Mounia Khelfaoui. 2020: “Population density, a factor in the spread of COVID-19 in Algeria: statistic study”

Arunava Bhadra ·Arindam Mukherjee ·Kabita Sarkar2. 2020: “Impact of population density on Covid-19 infected and mortality rate in India.”

Hamit Coskun. 2020: “The spread of COVID-19 virus through population density and wind in Turkey cities”

Christopher Dye. 2020: “The scale and dynamics of COVID-19 epidemics across Europe”

W.Cao. 2020: “Important factors affecting COVID-19 transmission and fatality in metropolises”

DARRYL COHEN. 2015: “Understanding Population Density”

ManuelFebrero-Bande: “Measures of influence for the functional linear model with scalar response”

https://www.who.int/health-topics/coronavirus#tab=tab_1