Geovid: Spatial Analysis & GIS of Coronovirus-2019 in China

Ivy HAU Jia Yi
Singapore Management University
ivy.hau.2018@sis.smu.edu.sg

Jasky ONG Qing Hao
Singapore Management University
jasky.ong.2017@sis.smu.edu.sg

Shermin TAN
Singapore Management University
shermin.tan.2018@sis.smu.edu.sg

ABSTRACT
This is the abstract. It consists of two paragraphs.

1. INTRODUCTION
Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness.

The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes. With such similar symptoms to a common "cold", this virus outbreak has spread to more than 200 countries and regions, affecting more than 60,000,000 people around the world, with more than 50% of recovery rate. In addition, it has also swept out more than 1,400,000 deaths, as of November 2020.

Presently, there are no vaccines available globally to protect against this virus, and the best way to prevent and slow down transmission is to be well informed about the COVID-19 virus, the disease it causes and how it spreads. In due of that, it will be significant to analyse the spread of COVID-19, understand where, when, and how it occurred by studying the spatio-temporal patterns of confirmed, recovered and death cases.

2. MOTIVATION & OBJECTIVES
In due of the limited resources of COVID-19 data Singapore has to offer, we have decided to shift our focus to China where the first COVID-19 case was identified in November 2019. We've acquired data from Harvard Dataverse featuring daily updates from the period of January to September 2020 of COVID-19 China confirmed, recovered, dead and foreign* cases. Together with these data, we will be able to do an in-depth analysis of the spatio-temporal patterns of the disease spread and understanding it could greatly improve and deter the size of this threat.

Furthermore, we aim to provide a web-based geospatial application with a web-enabled geospatial analytical tool to identify areas with COVID-19-related cases and pinpoint the factors affecting the spread. Particularly, this project will focus on the following objectives:

1. Visualize amount/density of the cases by performing relevant exploratory data analysis;
2. Conduct global and local spatial analysis to uncover spatial correlation patterns and its influencing factors;
3. Create a web-based interface through R Shiny with relevant user inputs to display relevant data analysis; The application will comprise(s) of the following analysis requirements:
    - Graphical/Geographical visualization framework that can display the amount/density of the COVID-19 cases over time;
    - Map visualization framework that supports macro and micro views;
    - Customizable analysis based on-demand through user's input.

On the whole, we can evaluate results based on the findings, better plan and provide recommendations to further mitigate its spread.

*Foreign cases will not be analyzed in this paper as we are only focusing on the spread towards the country and its residents.

3. RELATED WORK
Multiple studies regarding COVID-19 in China have been conducted after the first outbreak in Wuhan City. A study studied the spatial autocorrelation and factors of COVID-19 on a provincial level using methods such as Global Moran I, Local Moran I, Gertis-Ord Gi*.

The Global Moran I, will be used to measure the spatial characteristics of cases in the entire region, analysing the overall spatial correlation and differences between the regions. If Moran's I statistics shows a positive correlation, it implies that there is spatial clustering. Likewise, a negative correlation will imply a checkerboard spatial pattern. Local Moran's I was later used to study correlation patterns between regions. Gertis-Ord Gi* then identifies the spatial association between the hot and cold spots that is statistically significance. These statistical methods are relevant and should be incorporated to our GEOVID application.

Another similar study was done on the spatial and temporal differentiation of Covid-19 in China and its influencing factors. Like the previous study, uses Global Moran I, Spatial Weight Matrix and Local Moran's I. These statistical methods were also used to study the spatial correlation in China and its regions.

Based on these two research papers, Global Moran I and Local Moran I were used to analyse the spatial correlation of Covid-19 in China. These analysis methods will give a guide of relevant statistics methods for our GEOVID application.

4. RESEARCH METHODS

As previously mentioned, Global Moran I and Local Moran I are statistical methods often used to study the spatial correlation regarding Covid-19 cases. As such, for the GEOVID application, both Global and Local Moran will be incorporated it the analysis. In addition, spatial weight matrix will also be used in the Local Moran statistical Analysis to define the neighbour.

4.1 Global Moran's I

Global Moran I is a statistical method used to calculate the relationship between factors to its surrounding area. It evaluates if patterns are clustered, dispersed, or random. It also calculates the x score and p-value to find out the significance of index. P-value and z-score will prove is the null hypothesis is statistically significant. If the p-value is small and z-score falls outside the confidence level, null hypothesis will be rejected, proving that spatial patterns is statistically significant. The index value will indicate signs of spatial patterns, specifically, index near +1 would indicate clustering pattern while −1 would indicate dispersion.

The formula for Global Moran's I is as followed:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

In this study, Global Moran will be used to measure the spatial correlation of Covid-19 rates per 10,000 people, to the whole of China. Moran.test() and monte carlo simulation will be done to determine the significance of spatial patterns present.

4.2 Local Moran's I

Local Moran I is a local spatial autocorrelation analysis method based of Global Moran I. Local Moran I study the relationship between region, in this case, between cities in China. It also identifies the spatial outliers. Positive Local Moran index would suggest that region is surrounded with similar values, which will be identified as a cluster, while a negative index would suggest that region is surrounded by dissimilar values, which could be identified as an outlier.

The formula for Local Moran's I is as followed:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}(x_i - \bar{X})$$

In this study, Local Moran's I will be used to measure the local spatial correlation of Covid-19 rates per 10,000 people to cities in China.

5. APPROACH

5.1 Data Collection

Data of the coronavirus virus spread in China (January to September, 2020): Confirmed, Recovered and Death cases, Population Data (2010)* in Tab-Delimited format, China City/Province basemaps in Shapefile (SHP) format, are obtained from Harvard Dataverse: China COVID-19 Daily Cases with Basemap.

*Kindly take note that population data after year 2010 of each city are not readily available online as most of the sources only provide population data based on province or as an overall count instead. Therefore, we will make use of the population data of year 2010 given in the dataset above.

5.2 Data Cleaning and Wrangling

For all the respective data: Confirmed, Recovered and Death cases, we will only need to extract the daily count based on the date that the data was recorded. We will be focusing only from January to September 2020 as these datasets are the most complete range in the originals. In addition, duplicates will also need to be checked and removed to prevent skewed data analysis. Standardization of data, such as, ensuring columns of each data files and their data types are the same before reclassifying selected columns, is conducted. Consequently, it will then produce an unbiased comparison and analysis.

5.3 Data Transformation

6. SYSTEM ARCHITECTURE

Our application is built using R programming language, a widely used language among statisticians and data miners for developing statistical software and data analysis. We've constructed it upon an integrated development environment (IDE), Rstudio application, due to its extensive availability of R packages that we could import and deliver ease in developing our system architecture. Moreover, our application will be deployed to Shiny.io for further user interactions through their own private browser.

R Packages used for application development as below:

———————- | ———————- sp | sf | spdep spatstat | tidyverse | tidygeocoder lubridate | leaflet | classInt rgdal | raster | readxl rgeos | rsconnect | maptools dplyr | ggplot2 | ggpubr tmap | shiny | shinythemes

7. APPLICATION

8. LIMITATIONS

9. DISCUSSION

10. FUTURE WORK

11. ACKNOWLEDGEMENT

The authors would like express our special thanks of gratitude to Dr. Kam Tin Seong, Associate Professor of Information Systems at School of Information Systems, Singapore Management University, for giving us the opportunity to take part in this research project and providing kind support and guidance throughout this journey.

12. REFERENCES