Lab Report

Title: Prospectus to final project: spatial epidemiology of the Coronavirus in the US
Notice: Dr. Bryan Runck
Author: Mohsen Ahmadkhani
Date: 9/29/2021

Project Repository: https://github.com/mohsen-gis/GIS5571.git
Google Drive Link: -
Time Spent: 4 hours

Abstract

During the past two years, the COVID-19 pandemic has spread to almost all countries around the world and all the US states. The number of confirmed cases has been rapidly increasing before the vaccination program starts to over 53.7 million along with 1.3 million deaths as of November 15 globally. The pandemic has also imposed an unprecedented economic burden on every country in the world. Estimations show a rate of up to a -6% fall in the global economic growth in 2020 (Congressional Research Service, 2020). It is also predicted that over 100 million people could experience extreme poverty as a result of this global economic contraction (Congressional Research Service, 2020). Therefore, it is crucial to study this unforeseen pandemic from different angles like geographical information systems (GIS). In the proposed research I will study the spatial epidemiology of the disease through spatial clustering and correlation analysis considering temperature as an environmental factor.

Problem Statement

During the past two years, the COVID-19 pandemic has spread to almost all countries around the world and all the US states. The number of confirmed cases has been rapidly increasing before the vaccination program starts to over 53.7 million along with 1.3 million deaths as of November 15 globally. The pandemic has also imposed an unprecedented economic burden on every country in the world. Estimations show a rate of up to a -6% fall in the global economic growth in 2020 (Congressional Research Service, 2020). Hence, it is crucial to understand the spatial behavior of the disease in the US.

Table 1. The list of required data sets for the proposed study.

#	Requirement	Defined As	(Spatial) Data	Attribute Data	Dataset	Preparation

1	US COVID-19 data	The stats of the disease within the US to date.	CSV file with county-level fips code	County name, fips, cases, deaths, date	NY Times	ETL and convert to GeoPandasDataframe
2	US county-level boundaries	The boundaries of all US counties	Polygon Shapefile	County name, state name, fips, geometry	US Census Bureau	Attr join to COVID data
3	US Temperature data	The monthly temperature of the US at county level	JSON file with county-level fips code	County name, fips, temperature, date	Google Earth Engine	ETL and convert to Pandas DF

Input Data

In the proposed research I with download three sets of data to accomplish the project. First, the COVID-19 counts and deaths data from NewYork Times github repository that is being updated daily. Second, US counties' boundaries as a shapefile downloadable from the US census bureau website. Third, aggregated temperature data for the US at county level monthly.

Table 2. Input data description.

#	**Title**	**Purpose in Analysis**	**Link to Source**
1	US COVID-19 data	Raw input dataset for performing spatial clustering analysis and correlatin analysis.	NYTimes GitHub Repo
2	US county-level boundaries	To run an attr join analysis with the Covid data to make it spatial	US Census Bureau
3	US Temperature data	To run the correlation analysis	Google EE

Methods

In this study, for implementing the entire pipeline I will use Python programming language. In the process, I will make use of ArcPy module to implement cluster analysis. I will use global Moran's I analysis to understand the spatial distribution of the disease cases in the country. It will be clustered, or dispersed, or random. Next, I will use local Moran's I and Getis-Ord clustering methods to identify all spatial clusters of the disease in the country. And finally, I will use regression analysis as well as Pearson correlation analysis to assess the possible correlation between the disease and temperature.

Results

The expected results are a number of clustered in highly populated areas across the US. Since I'm using the cumulative number of disease cases, the clusters are expected to appear around the big, populated cities like Manhattan, LA, and Miami. The result will be a set of polygons as a choropleth map with a color scheme highlighting the areas according to their type of cluster. The

clusters would be one of the following four groups: high-high, high-low, low-high, or low-low (figure 1). The clusters of interest would be those labeled as high-high and high-low meaning that a county with higher rate of the disease is surrounded a county with either high rates of the disease or low respectively. Also, the expected result of Pearson analysis should be a table with numbers indicating the correlation coefficients showing the importance of the correlation. The null hypothesis for this analysis is that the temperature has no correlation with the disease prevalence.

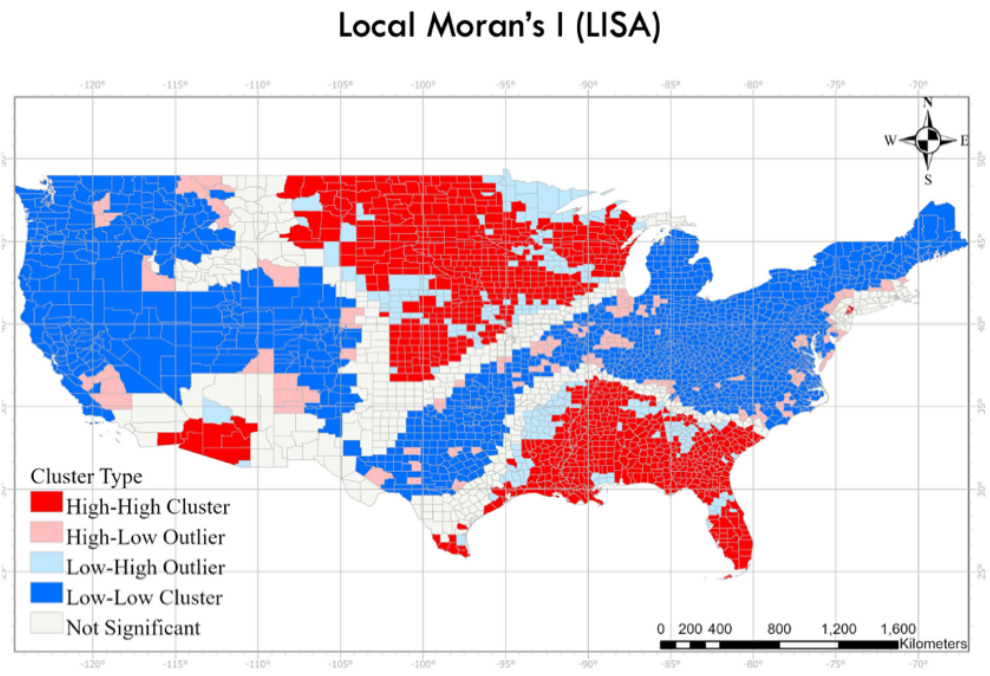

Figure 1: A sample output of the local Moran's I analysis for the US.

Results Verification
To evaluate the results of the processes I will compare my results with the articles that are peer reviewed and published in this area. This comparison will confirm or invalidate my results based on their consistency with the published literature. The potential reference literature are listed in the references section [1, 2 ,3, 4].

Discussion and Conclusion

Technical
After finalizing this project, I expect to be proficient with using ArcPy functions through a python script. I also anticipate being well familiar with the ETL process and using REST APIs from different websites to automatically download and manipulate freely accessible data. Also, this will be an illustration of implementing a pipeline for automatic process of a spatiotemporal analysis for a real-world problem at a continental scale.

Analytical
From an analytical point of view, my research will generate knowledge about the spatial and spatiotemporal behavior of a critical disease like COVID-19 pandemic. It will reveal whether an environmental factor like temperature has any impact on the prevalence of the disease or not. It also will confirm that the disease has a clustered spatial distribution in the US. This information will be potentially useful for the health care officials and the policymakers to revise and optimize the social restrictions accordingly and efficiently allocate resources across the country.

References

1. Kang, D., Choi, H., Kim, J. H., & Choi, J. (2020). Spatial epidemic dynamics of the COVID-19 outbreak in China. *International Journal of Infectious Diseases*, *94*, 96-102.
2. Vahabi, N., Salehi, M., Duarte, J. D., Mollalo, A., & Michailidis, G. (2021). County-level longitudinal clustering of COVID-19 mortality to incidence ratio in the United States. *Scientific reports*, *11*(1), 1-22.
3. Bilal, U., Tabb, L. P., Barber, S., & Diez Roux, A. V. (2021). Spatial Inequities in COVID-19 Testing, Positivity, Confirmed Cases, and Mortality in 3 US Cities: An Ecological Study. *Annals of internal medicine*.
4. Andersen, L. M., Harden, S. R., Sugg, M. M., Runkle, J. D., & Lundquist, T. E. (2021). Analyzing the spatial determinants of local Covid-19 transmission in the United States. *Science of the Total Environment*, *754*, 142396.

Self-score

Category	Description	Points Possible	Score
Structural Elements	All elements of a lab report are included **(2 points each)**: Title, Notice: Dr. Bryan Runck, Author, Project Repository, Date, Abstract, Problem Statement, Input Data w/ tables, Methods w/ Data, Flow Diagrams, Results, Results Verification, Discussion and Conclusion, References in common format, Self-score	28	**28**

Clarity of Content	Each element above is executed at a professional level so that someone can understand the goal, data, methods, results, and their validity and implications in a 5 minute reading at a cursory-level, and in a 30 minute meeting at a deep level **(12 points)**. There is a clear connection from data to results to discussion and conclusion **(12 points)**.	24	**24**
Reproducibility	Results are completely reproducible by someone with basic GIS training. There is no ambiguity in data flow or rationale for data operations. Every step is documented and justified.	28	**28**
Verification	Results are correct in that they have been verified in comparison to some standard. The standard is clearly stated **(10 points)**, the method of comparison is clearly stated **(5 points)**, and the result of verification is clearly stated **(5 points)**.	20	**20**
		100	**100**