Lab Report

Title: Draft 1
Notice: Dr. Bryan Runck
Author: Mohsen Ahmadkhani
Date: 11/292021

Project Repository: *https://github.com/mohsen-gis/GIS5571.git*
Google Drive Link: *<if applicable with data, notebooks, etc.>*
Time Spent: *<report to the nearest quarter hour>*

Abstract

250 words max. Clearly summarize the following major sections. Each gets one or two sentences.

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. To study the spatial epidemiology of the COVID-19 pandemic it is extremely crucial to learn about the spatial and spatiotemporal clusters of the disease in a monthly resolution. This analysis will expand our knowledge in terms of finding the counties in the US that are having a critical situation regarding the spread of the disease. In addition, assessing the correlation between the disease cases' distribution and the temperature as an environmental factor would let the policymakers know more about the nature of the disease.

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 this research I have downloaded 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 used Python programming language and PostgreSQL. In the process, I used ArcGIS Pro to implement a monthly cluster analysis. I used global Moran's I analysis to understand the spatial distribution of the disease cases in the country. Next, I applied local Moran's I clustering method to identify all spatial clusters of the disease in the country on a monthly basis. The remaining part is using a correlation analysis to assess the possible correlation between the disease and temperature.
Some of the SQL scripts I used for the data preparation are as follows (the complete codes and scripts will be pushed to my repository):

```
population us county:

https://www.openintro.org/data/?data=county_complete

https://www.openintro.org/data/csv/county_complete.csv


Create table us_county_pop and import pop csv to the table
-- Table: public.us_county_pop

-- DROP TABLE public.us_county_pop;

CREATE TABLE public.us_county_pop
(
    fips character varying(10) COLLATE pg_catalog."default",
    state_name character varying(100) COLLATE pg_catalog."default",
    county_name character varying(100) COLLATE pg_catalog."default",
    pop2017 character varying(15) COLLATE pg_catalog."default"
)

TABLESPACE pg_default;

ALTER TABLE public.us_county_pop
    OWNER to postgres;


# zero padding to the fips that are miising an initial 0

UPDATE us_county_pop
SET fips = concat(0,fips)
where length(fips)=4

# create a final county_pop table as follows:

create table county_pop as(
select uc.geoid, uc.county_name, uc.state_name, ucp.pop2017, uc.geom
from us_county as uc left join us_county_pop as ucp on uc.geoid = ucp.fips
)
```

```
create table daily_time_series as(
with dataset as     (
 select
    geoid,
    county_name,
    state_name,
    date,
    cases,
    deaths,
    pop2017 as population

from
    time_series
order by
    geoid,
    date     ), timelag as
    (
select
    geoid,
    county_name,
    state_name,
    date,
    population,
    cases as total_cases,
    cases-lag(cases,1)  over w as new_cases,
    deaths as total_deaths,
    deaths-lag(deaths,1)  over w as new_deaths
from
    dataset

window w AS ( PARTITION BY
    geoid
order BY
    date )     )
select
    geoid,
    county_name,
    state_name,
    date,
    population,
    total_cases,
    coalesce(new_cases,0) as new_cases,
    round((coalesce(new_cases,0)*10000)/population::numeric,4) as daily_incidence,
    round((total_cases*10000)/population::numeric,4) as cum_incidence ,
    total_deaths,
    coalesce(new_deaths,0) as new_deaths,
    round(coalesce((new_deaths*100/total_cases),0)::numeric,4) as daily_death_rate,
    round((total_deaths*100/total_cases)::numeric,4) as cum_death_rate

from
    timelag
where
    total_cases >0

    )
```

Results

At this point, the cluster analysis has been accomplished and the purely spatial and space-time clusters have been identified for the monthly data. After aggregating the data using PostgreSQL, I performed the local Moran's I and generated an animation depicting the monthly trend of disease cluster formation in the US (see Figure 1). The results of Retrospective Space-Time permutation scan statistics analysis (Figure 2) showed that there are 5 space time clusters in the US. Ignoring the one huge cluster in the Midwest, there has been a small cluster in Missouri in June 2021. Another cluster is the one covering some parts of Ohio, Pennsylvania, Michigan, and New York states. This cluster has been started in October 2021 and is still ongoing. In the western US, there has been a cluster during May and June 2021 covering parts of Oregon and Washington states. And the last cluster has happened in the southern US, covering Florida, Alabama, and some parts of their neighboring states. This cluster started in July 2021 and ended in September 2021. The last part of the analysis is the correlation analysis that is still not performed.

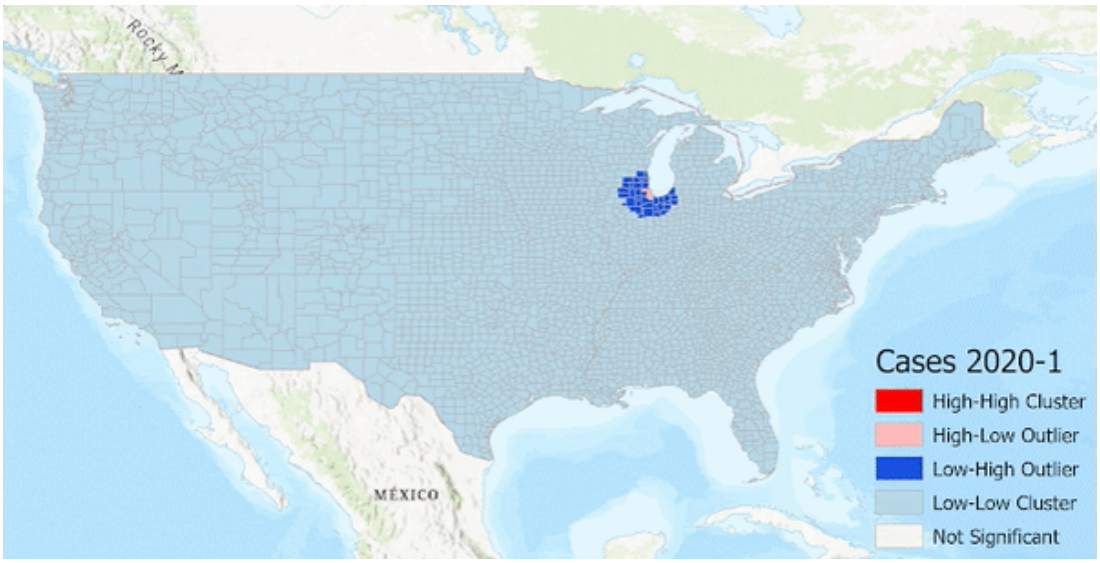

Figure 1. The result of monthly cluster analysis for the contiguous US.

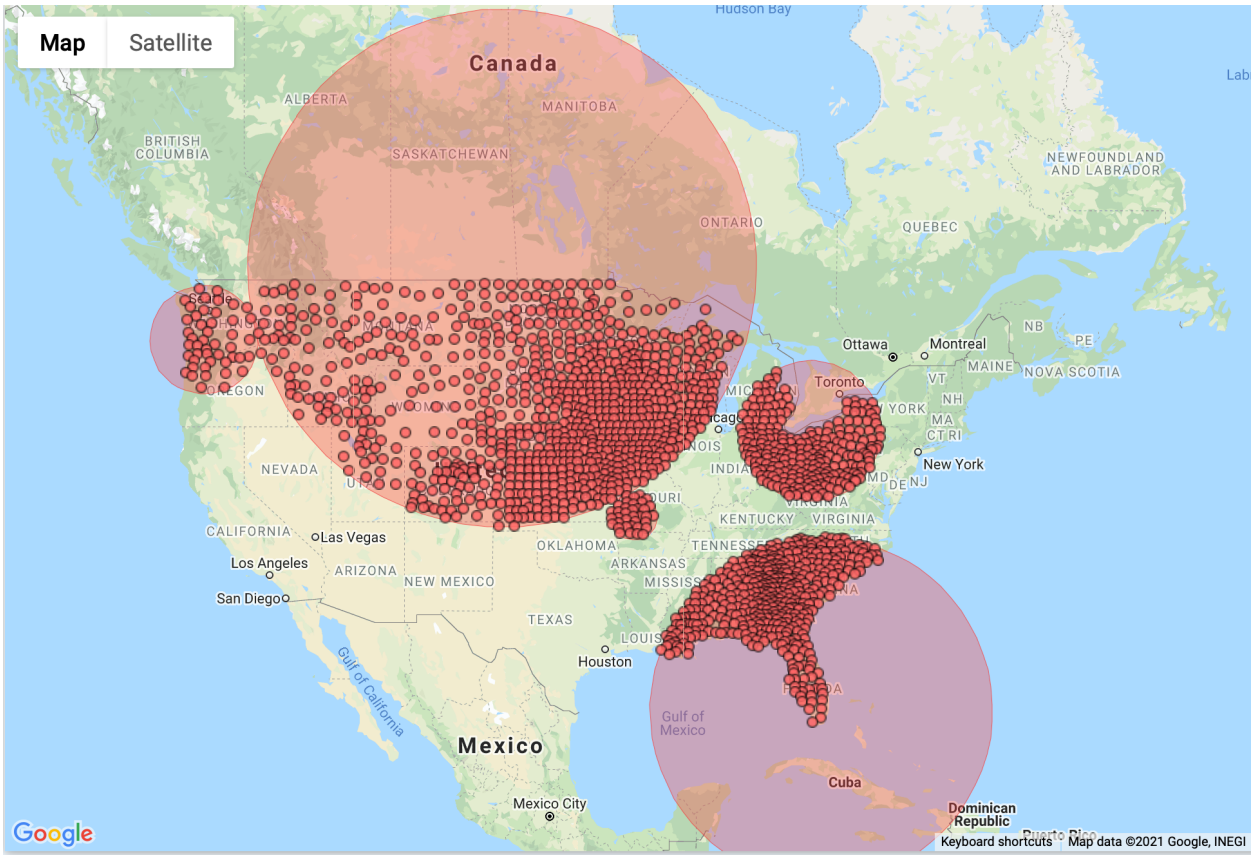

Figure 2. The result of Retrospective Space-Time permutation scan statistics analysis.

Results Verification

The verification for the analyses that are performed in this research have all been done using the P-values and the Z-scores. To evaluate the result of global Moran's I, I used Z-score to verify the significance of the results. To evaluate the results of the local Moran's I, I used p-values as an indicator of the significance of the detected clusters. The threshold for the evaluation of the results have been 95 percent or p-value less than 0.05. The cluster analysis was done using monte Carlo simulation for 999 times of iteration.

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	**24**
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	**20**
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	**21**
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	**85**