GEOG788P Models and Methods for Spatial Data Science
Project Proposal
Guimin Zhu

Introduction

COVID-19 is the ongoing global pandemic that changed our daily life so dramatically. I have been participating in an NSF project on the research of mobility and COVID-19 in Florida. Thus, I think it would be a good idea to investigate Maryland (although Maryland is not a COVID hotspot). The public, especially Maryland residents, would like to see some interesting findings of how Maryland corresponded in this pandemic.

Research Questions

As introduced in the first section, the COVID-19 in Maryland is a potentially beneficial topic for the public. Thus, my initial research questions are:

QUESTION 1: What are the spatiotemporal patterns of the COVID-19 pandemic in Maryland in terms of positive cases, the volume of testing, the number of death, etc.?

The spatiotemporal patterns of the COVID-19 in Maryland will be investigated. The spatial distributions of the COVID-19 measures, e.g. the hot spots and cold spots, will be mapped. The spread of the COVID-19 over time will be characterized as well. As for the measures, potential data include positive cases, the volume of testing, the number of death, etc. A subset of these measures (or other derived measures, e.g. the incidence rate, which is the total case number by the census tract population) will be selected according to the scope and time frame of this project.

QUESTION 2: What are the demographic and socioeconomic determinants that are impacting the COVID-19 outbreak in Maryland and how these factors are impacting COVID-19?

A Geographic Weighted Regression model and a machine learning regression model will be applied to investigate the demographic (e.g. age, race, and ethnicity) and socioeconomic (e.g. median household income, median household size, and education) factors that are impacting the spread of the COVID-19 pandemic. Another question of how these factors are impacting it will also be studied.

Study Area and Sources of Data

The study area of this project is Maryland with the census tracts as the spatial units. There is county level research and I think a more refined level (census tract level) would provide more interesting findings. The detailed data description and data sources are listed below in Table 1.

Table 1. Data sources

Data	Detail	Type	Source
Maryland COVID-19	COVID-19 positive cases, testing, and death numbers at zip code level	csv	Maryland Department of Health https://coronavirus.maryland.gov/
Maryland political boundaries	State, county, zip code, and census tract boundaries	shapefile	Maryland's Open Data Portal https://opendata.maryland.gov/
Maryland health data	Localtion of hospitals, etc.	shapefile	Maryland's GIS Data Catalog https://data.imap.maryland.gov/
Maryland demographic data	Race, ethnicity, education, income, etc.	csv	2018 American Community Survey https://www.census.gov/programs-surveys/acs

Python Tools

A variaty of Python tools and modules are expected to be used in this project in order to accomplish the functionalities of data processing, geo-visualization, statistical and spatial analysis, and regression. Some potential Python tools include: Pandas, GeoPandas, Matplotlib, seaborn, SciPy, scikit-learn, and PySAL. In addition, ArcGIS Pro by ESRI will be used to conduct the Geographic Weighted Regression.

Envisioned Challenges

During data processing step, all the data from different data sources are different in terms of spatial units, data quality, etc. They can be either quantitative or qualitative. It is foreseen that the data are so messy that much effort will be required for data pre-processing. Running the machine learning regression model is another major challenge because the tuning process for optimized model performance is relatively difficult. Finally, after I get the model, how to interpret the model (i.e. conclusions on what the model actually tells) is another challenge.