GEOG788P

Final Project Paper

Visualization of COVID-19 Spread in Maryland and Geo-spatial Modelling the Spatiotemporal and Socioeconomic Impacting Factors

Guimin Zhu

1. Introduction

Starting from January 20, 2020, when the first confirmed case of the 2019 coronavirus disease (COVID-19) in the United States was reported, the United Stated has the most confirmed case number and death number in the world (over 13.8 million and over 270 thousand respectively, as of December 1, 2020). Although the WHO declared COVID-19 a global pandemic on March 11, 2020, the COVID-19 situation in the United States did not get under control. The third wave indicated by public health experts has already arrived in November, with over 200 thousand daily new cases on November 27. Following the declaration of national emergency, states started issuing stay-at-home orders, implementing travel and mobility restrictions to contain and combat the spread of COVID-19 by limiting individuals' movements and personal interactions outside of their home to only those necessary to obtain or provide essential services or conduct essential activities. Research shows that mobility restrictions and interventions helped fight COVID-19 (Espinoza, Castillo-Chavez, & Perrings, 2020) by reducing the COVID-19 spread (Gao et al., 2020).

Meanwhile, Maryland, although not one of the COVID-19 hotspots, has over 215 thousand confirmed cases and over 4.8 thousand deaths as of December 1, 2020. Living in Maryland, the public is interested in how the COVID-19 pandemic has evolved, how the Maryland Department of Health reacted and mitigated the pandemic, when the second/third wave would come, etc.

In this study, I examine what is the spatiotemporal pattern of the COVID-19 pandemic in Maryland in terms of positive cases, e.g. where the hot spots and cold spots are. I also analyzed to determine what are the demographic and socioeconomic determinants that are impacting the COVID-19 outbreak. How did these factors change over the 2 time periods (March 15 to June 30 and July 1 to August 31)? Details about how these 2 time periods are obtained will be introduced in Section 3.

2. Related Work

There has been a number of studies related to COVID-19 and its socioeconomic impacts published this year since the outbreak of the pandemic, both in the U.S. and overseas. Research on the effects of social distancing restrictions examines the association between the rate of human mobility changes (i.e. change rates of the median travel distance and the median hoe dwell time of mobile phone users) and the rate of confirmed COVID-19 cases (Gao et al., 2020). Results show that stay-at-home orders were associated with the reduction in COVID-19 spread.

Currently, the GIS scientists have published their studies on COVID-19 on multiple scales and using several techniques. Chen, Jiao, Bai, & Lindquist (2020) conducted OLS and GWR models to explore and model the spatial factors of COVID-19 in New York City, one of the hot spots in the early pandemic. Research using GWR and MGWR with 25 environmental, socioeconomic, topographic, and demographic variables to investigate and predict COVID-19 has been carried at the county level all over the continent U.S. (Mollalo, Vahedi, & Rivera, 2020). Other machine learning algorithms, e.g. Artificial Neural Network, have been applied in the prediction of COVID-19 incidence rate nationwide with 57 candidate explanatory variables (Mollalo, Rivera, & Vahedi, 2020).

Random Forest (RF) regression model is one of the commonly used machine learning algorithms for geo-spatial data, and the researchers developed the Geographical Random Forest, which extends the ability to address spatial heterogeneity than the normal RF models (Georganos et al., 2019).

3. Data and Methods

3.1. Data Preparation

There are mainly 2 domains of data: COVID-19 related data, which are obtained from Maryland Department of Health, and Maryland demographical and socioeconomic data, which are obtained from the 2018 American Community Survey (ACS). The COVID-19 related data include positive cases and confirmed deaths by county, age, gender, race, and ethnicity, which are from March 15 to August 31. The only available sub-county level data are positive case numbers, which are from April 11 to August 31. In addition, daily testing volume and total population tested data are downloaded, but not used in this study. Future work may find these data useful. Maryland statistics are downloaded from 2018 ACS, with categories including population, age, gender, race and ethnicity, income and poverty, housing, education, and employment at the census tract level. The above data will be cleaned and used in the regression models.

Other data, such as Maryland political boundaries, are obtained from Maryland's Open Data Portal and Maryland's GIS Data Catalog. Table 1 below has detailed descriptions of the data.

Table 1. Data used in this study

Data	**Description**	**Source**
COVID-19 related data	Cases by county, age, gender, race and ethnicity & cases by zip code	Maryland Department of Health
Maryland boundaries	County-level, zip code level, and census tract level	Maryland's Open Data Portal and Maryland's GIS Data Catalog
Maryland statistics	Populations, age, gender, race/ethnicity, income and poverty, housing, education, and employment	2018 ACS
Maryland hospital locations	Hospitals licensed by Maryland DOH	Maryland's GIS Data Catalog

3.2. Data Processing and Geo-visualization

First, I visualize the trend of COVID-19 cases at the zip code level, both the cumulative case number and daily new case number (Figure 1).

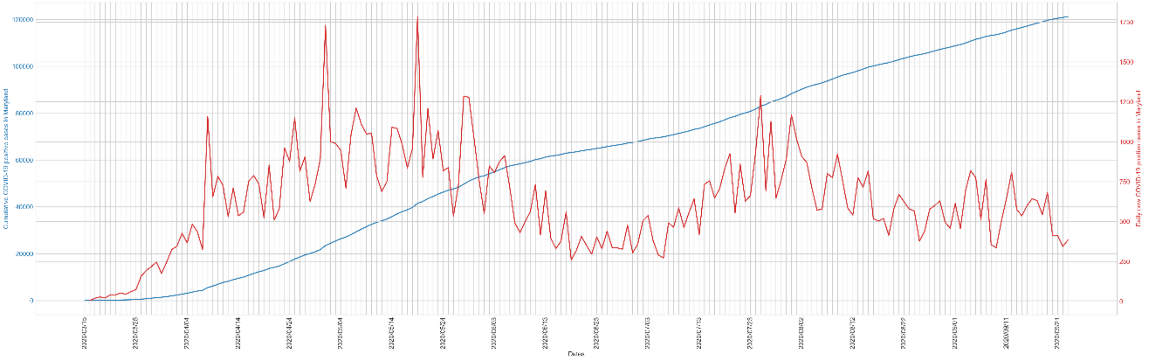


Figure 1. Maryland cumulative COVID-19 positive cases (blue) and daily new cases (red) from March 15 to August 31.

From Figure 1, we can easily identify 2 time periods based on the daily new cases: March 15 to June 30 and July 1 to August 31. We assume that these 2 time periods will have different temporal effects, which leads to separate spatial modeling. The flat curve of the cumulative case number during the mid-June to mid-July period indicates low increases, but after that, the situation seems to be getting worse, which suggests that the social distancing restrictions were effective to constrain the COVID-19 spread until the stay-at-home order was lifted and phased reopening was implemented.

Then, Figure 2 below shows the cumulative COVID-19 case number rankings by county as of August 31. We can see that Prince George's County, Montgomery County, and the Baltimore area (Baltimore City and Baltimore County) have most of the cases. These areas are the focus of this study.

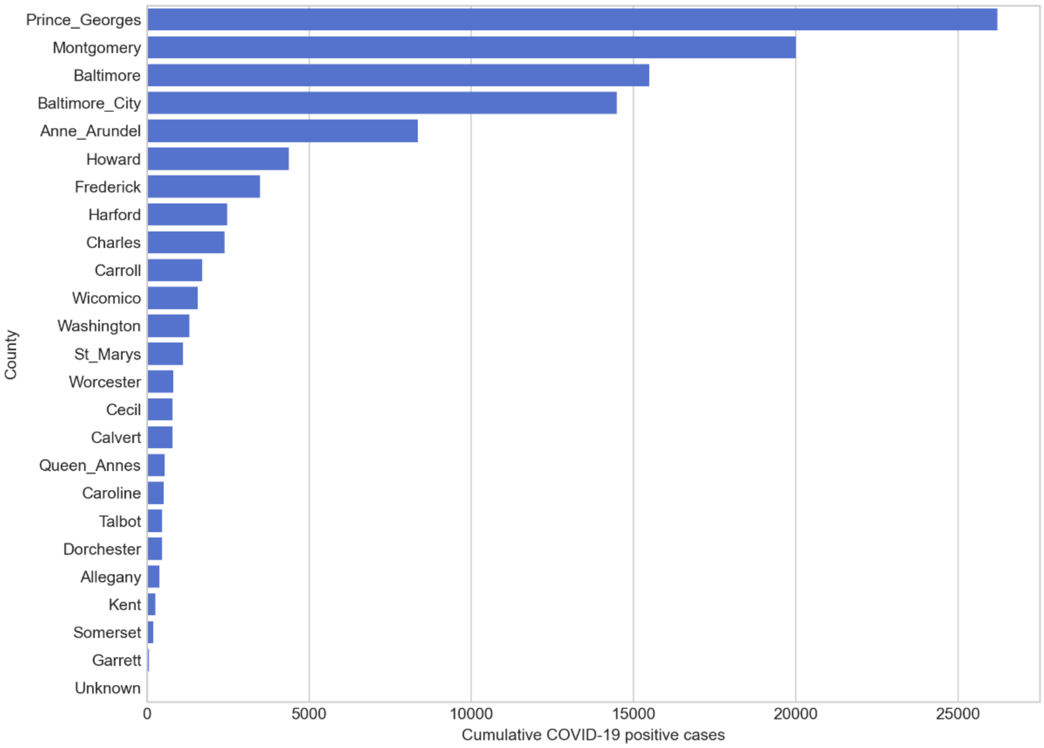


Figure 2. Maryland cumulative COVID-19 positive case number rankings by county (as of August 31).

To unify the spatial units, I crosswalk the COVDI-19 case data from zip code level to census tract level. The zip code level case data are multiplied by the ratio of census tract population to zip code population, and then census tract level data are aggregated.

3.3. Spatial Clustering

To identify the COVID-19 hot spots and cold spots, I perform the hotspot analysis using Local Moran's I to examine the clusters of COVD-19 cases. With hot spots and cold spots, state health experts and officials can have a better understanding of where the

COVID-19 situation needs more focus. The results can be greatly useful for policymaking.

Regionalization is a technique to aggregate the basic spatial units into larger regions. In order to investigate the regions of COVID-19 case numbers and to better adjust the medical resources, I also run a regionalization analysis (Duque, Ramos, & Suriñach, 2007) using K-Nearest Neighbor (k=5, and 4 nearest neighbors).

3.4. Ordinary Least Squares and Random Forest Model

To examine the demographic and socioeconomic determinants that are impacting the COVID-19 spread, I run regression models on the COVID-19 case numbers. Three models are to be used in this study: Ordinary Least Squares (OLS), Random Forest (RF), and Geographically Weighted Regression (GWR). The OLS model and RF model are universal models, while the GWR model is a local model that considers the spatial non-stationarity (Oshan, Li, Kang, Wolf, & Stewart Fotheringham, 2019).

The dataset preparation is done in Section 3.1., and the selected explanatory variables include (1) population density, (2) race and ethnicity (White, Black, Asian, Native, Other), (3) age groups of 20 years interval (0 to 19, 20 to 39, 40 to 59, 60 to 79, 80 and above), (4) population with Bachelor degree or higher, (5) population below poverty level, (6) median household income, (7) unemployment rate, and (8) uninsured population. Obviously, the dependent variable is the new COVID-19 positive case number from March 15 to June 30 period and July 1 to August 31 period.

In order to perform the OLS model and compared the coefficients of top impacting variables, the whole dataset is standardized before fitting the model. To standardize the dataset, each variable subtracts the mean of the variable and then is divided by the standard error of the variable, leading to a variable distribution with the mean being 0 and standard error being 1.

An RF model is trained using the same dataset. The basic processing includes filtering out the extreme records and splitting them into 75% of the training set and 25% of the testing set. A python package, Scikit-Learn, is used for parameter tuning with 5-fold cross-validation. In the case of overfitting, some techniques, e.g. regularization, oversampling, stratify splitting, etc., will be used.

3.5. Geographically Weighted Regression

To better target the spatial context of COVID-19 case number determinants, while OLS and RF regression only model global processes, GWR takes the spatial heterogeneity into modeling. A previous case study on the obesity rates in the Phoenix metropolitan area shows the general procedure of GWR (Oshan, Smith, & Fotheringham, 2020). The GWR model is trained using the top features from the RF model results. Then the surfaces of parameters for the 2 time periods are compared to examine how these factors changed over the 2 waves.

4. Results

4.1. COVID-19 Trend

The cumulative COVID-19 case number and daily new case number have been visualized in Figure 1 and Figure 2. Then the distribution of cumulative case number at the zip code level is visualized in a time series, starting from April 12 and using 14 days as the interval. In the presentation, a GIF animation is provided, while here, the dates of April 12, June 21, and August 30 are selected. See Figure 3 below.

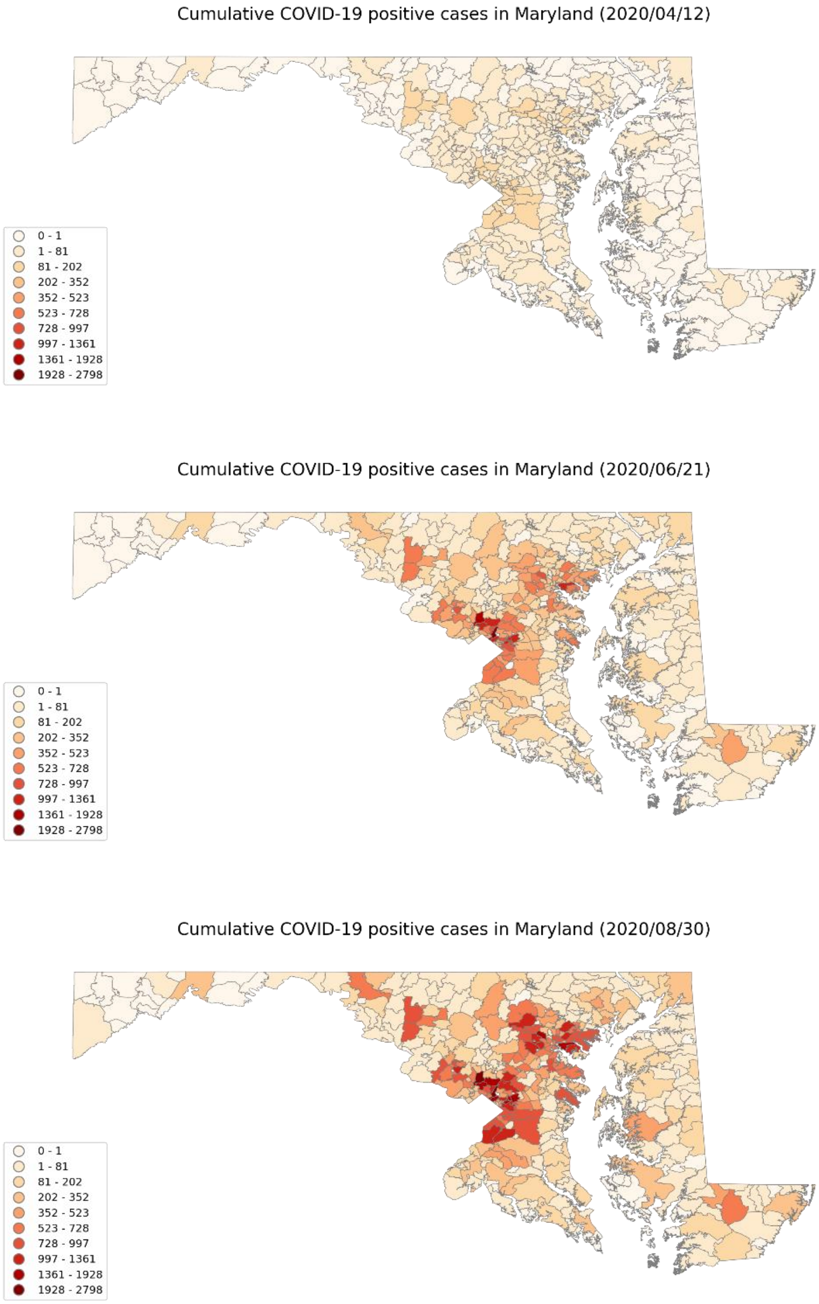


Figure 3. Cumulative COVID-19 positive cases in Maryland at zip code level (April 12, June 21, and August 30).

The COVID-19 evolving trend indicates a few hot spots in Maryland, especially in August. We can clearly identify the clusters near DC and the clusters near Baltimore.

4.2. Hotspot Analysis

First, the histograms of the 2 study time periods are visualized in Figure 4. A visualization of the spatial distribution is shown in Figure 5.

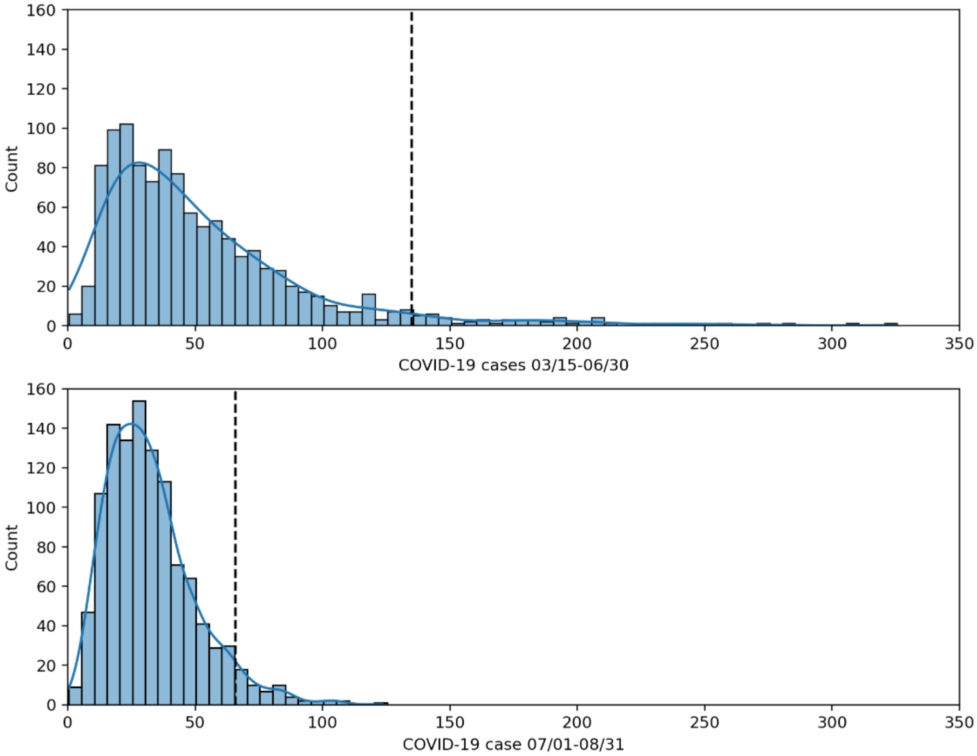
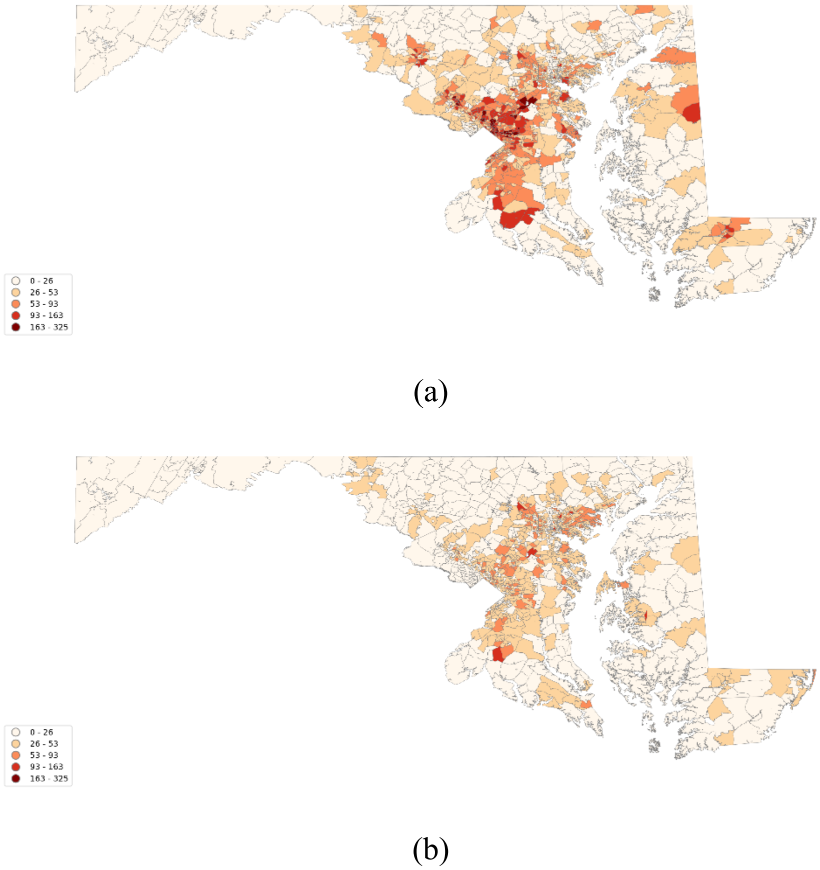


Figure 4. Histograms of new COVID-19 cases for the 2 time periods (March 15 to June 30 and July 1 to August 31).

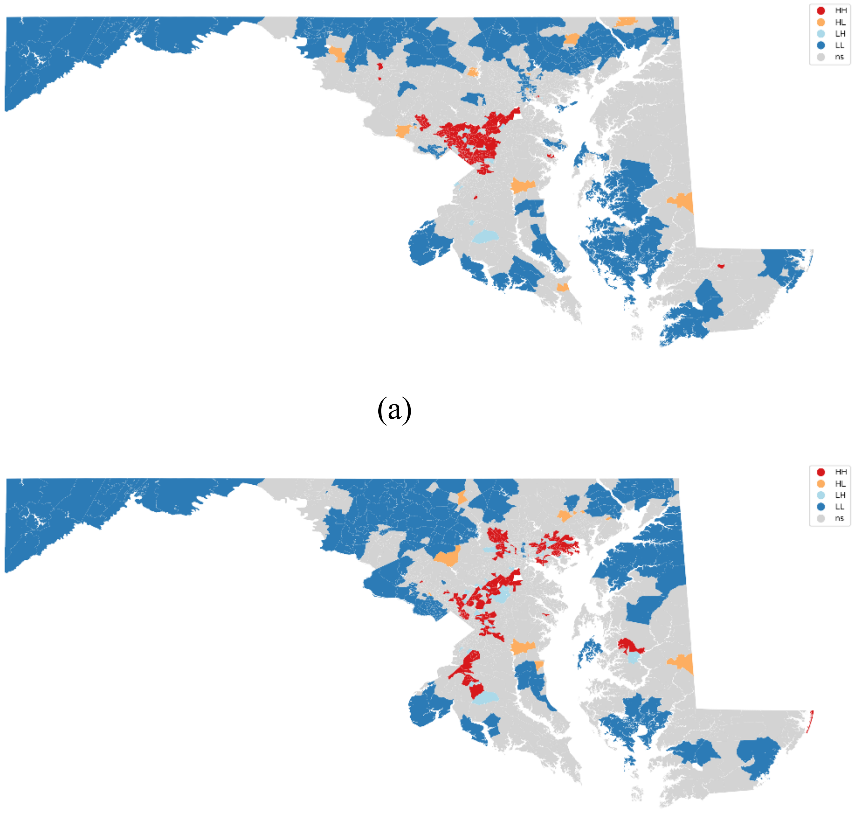


(a)

(b)

Figure 5. New COVID-19 positive cases in (a) March 15 to June 30, and (b) July 1 to August 31.

Overall, the first time period has more cases since the histogram shows a higher range of distribution, while the case numbers cluster in the low value ranges in the second time period. From Figure 5, we can see that the second time period has fewer cases and fewer hot spots. Based on these preliminary analyses, I conduct hotspot analysis using Local Moran's I to detect spatial hot spots and cold spots. The results are shown below in Figure 6. For the first time period, there is one big hot spot in the near DC to Ellicott City area, while in the second time period, there are about 5 separate locations of hot spots, which indicates the COVID-19 has been spread out.



(a)

(b)

Figure 6. Hotspot analysis for new COVID-19 positive cases in (a) March 15 to June 30, and (b) July 1 to August 31.

4.3. OLS Model and RF Model

After running the OLS model with the standardized dataset, the R^2 and the Adjusted R^2 are the model performance measurements. Looking at the variable coefficients (significant variables with the p-value greater than 0.05), the top impacting variables, both positive impacts, and negative impacts, are selected. See Table 2 down below.

Table 2. OLS model performance and the top impacting variables.

		03/15 – 06/30	**07/01 – 08/31**
R^2		0.4715	0.3840
Adjusted R^2		0.4635	0.3746
Top Variables	**1**	MHI (-)	MHI (-)
	2	Bachelor (+)	Unemployment (-)
	3	Other (+)	Age 20-39 (+)
	4	Unemployment (-)	Age 60-79 (+)
	5	Uninsured (+)	Bachelor (+)

Similar to OLS, the RF model also uses R^2 as the model performance evaluation metrics. In addition, Root Mean Square Error (RMSE) and MAE (Mean Absolute Error) are also applied to evaluate the errors. Details are in Table 3 below.

Table 3. RF model performance and the top impacting variables.

		03/15 – 06/30	**07/01 – 08/31**
R^2		0.4586	0.3893
RMSE		21.71	10.17
MAE		15.81	8.01
Top Variables	**1**	White	MHI
	2	Unemployment	Unemployment
	3	Other	Black
	4	MHI	White
	5	Black	Age 0-19

4.4. GWR Results

I select the top 6 variables that appear in both time periods as the features to fit the GWR model, which are (1) white population, (2) unemployment rate, (3) other population, (4) median household income, (5) black population, and (6) age group of 0 to 19. Using the python package, PySAL, the optimal search bandwidth for GWR is 114.0 and 104.0, respectively. The GWR model results are shown below in Table 4.

Table 4. GWR model performance.

	03/15 – 06/30	**07/01 – 08/31**
R^2	0.7323	0.6870
AIC	2496.21	2745.86

To compare the differences of the impacting features, the parameter surfaces are visualized using the same legend. The paired parameter surfaces can show the public health officials where the feature is impacting locally. For example, the white population surface maps are selected in Figure 7 below.

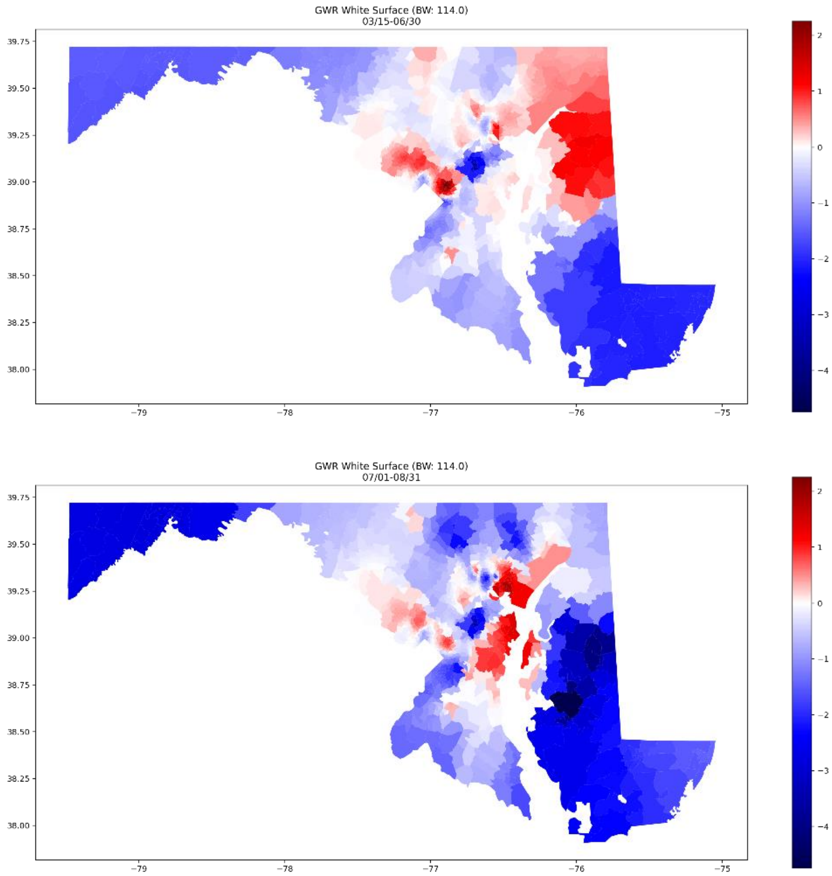


Figure 7. GWR parameter surface comparison for the white population.

We can see the obvious differences in the impact of the white population on COVID-19 cases in the 2 study time periods. In the first time period, the white population is significantly impacting the Silver Spring and Bethesda area and the border area with Delaware, while in the second time period, everything changed and the white population is impacting the Annapolis area and Baltimore area, where they are close to oceans and harbors.

5. Discussion and Conclusion

There are 2 waves of COVID-19 outbreak and the third wave has arrived in Maryland. Some counties, e.g. Prince George's County, Montgomery County, Baltimore City, and Baltimore county, have the most cases in Maryland, which should draw the attention of Maryland public health officials. The confirmed cases were clustered near the DC area (i.e. Silver Spring and Bethesda) in the first time period from March 15 to June 30, and then spread to several hot spots later in the second time period from July 1 to August 31.

The regression models, OLS model, RF model, and GWR model, show that some variables all ranked top in terms of the variable impacts, e.g. white population, black population, other population, median household income, unemployment rate, population with a bachelor degree or higher, etc., which meets the common sense and our expectations.

Among the 3 regression models, the GWR model has the best model performance (R^2 being 73.23% and 68.70% respectively). Also, the parameter surface maps show the localized variable importance, and public health officials and stakeholders may take these maps as references when making policies.

Some drawbacks of this study include the RF model overfitting issue. After applying some techniques, e.g. regularization, oversampling, stratify splitting, the overfitting has not been improved. Further studies on machine learning overfitting problem solutions are necessary. In addition, the results of regionalization are extremely unsatisfying due to unknown reasons. I will explore more in future work.

References

Chen, Y., Jiao, J., Bai, S., & Lindquist, J. (2020). Modeling the Spatial Factors of COVID-19 in New York City. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3606719

Duque, J. C., Ramos, R., & Suriñach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review*, *30*(3), 195–220. https://doi.org/10.1177/0160017607301605

Espinoza, B., Castillo-Chavez, C., & Perrings, C. (2020). Mobility Restrictions for the Control of Epidemics: When Do They Work? *SSRN Electronic Journal*, 1–14. https://doi.org/10.2139/ssrn.3496928

Gao, S., Rao, J., Kang, Y., Liang, Y., Kruse, J., Dopfer, D., … Patz, J. A. (2020). Association of Mobile Phone Location Data Indications of Travel and Stay-at-Home Mandates With COVID-19 Infection Rates in the US. *JAMA Network Open*, *3*(9), e2020485. https://doi.org/10.1001/jamanetworkopen.2020.20485

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., … Kalogirou, S. (2019). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, *0*(0), 1–16. https://doi.org/10.1080/10106049.2019.1595177

Mollalo, A., Rivera, K. M., & Vahedi, B. (2020). Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States.

International Journal of Environmental Research and Public Health, *17*(12), 1–13. https://doi.org/10.3390/ijerph17124204

Mollalo, A., Vahedi, B., & Rivera, K. M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of the Total Environment*, *728*(April), 138884. https://doi.org/10.1016/j.scitotenv.2020.138884

Oshan, T. M., Li, Z., Kang, W., Wolf, L. J., & Stewart Fotheringham, A. (2019). MGWR: A python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information*, *8*(6). https://doi.org/10.3390/ijgi8060269

Oshan, T. M., Smith, J. P., & Fotheringham, A. S. (2020). Targeting the spatial context of obesity determinants via multiscale geographically weighted regression. *International Journal of Health Geographics*, *19*(1). https://doi.org/10.1186/s12942-020-00204-6