Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

GIS-based spatial modeling of COVID-19 incidence rate in the continental United States

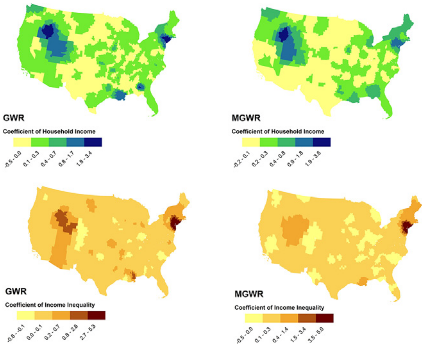
Abolfazl Mollalo [a],*, Behzad Vahedi [b], Kiara M. Rivera [a]

[a] *Department of Public Health and Prevention Sciences, School of Health Sciences, Baldwin Wallace University, Berea, OH, USA*
[b] *Department of Geography, University of California Santa Barbara (UCSB), Santa Barbara, CA, USA*

HIGHLIGHTS

- To explore relationship between 35 environmental, socioeconomic, and demographic variables and COVID-19 incidence in US
- Multiscale geographically weighted regression could explain 68.1% of the total variations of COVID-19 incidence in US
- Income inequality was an influential factor in explaining COVID-19 incidence particularly in the tri-state area

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 18 April 2020
Received in revised form 19 April 2020
Accepted 20 April 2020
Available online 22 April 2020

Keywords:
COVID-19
GIS
Multiscale GWR
Spatial non-stationarity

ABSTRACT

During the first 90 days of the COVID-19 outbreak in the United States, over 675,000 confirmed cases of the disease have been reported, posing unprecedented socioeconomic burden to the country. Due to inadequate research on geographic modeling of COVID-19, we investigated county-level variations of disease incidence across the continental United States. We compiled a geodatabase of 35 environmental, socioeconomic, topographic, and demographic variables that could explain the spatial variability of disease incidence. Further, we employed spatial lag and spatial error models to investigate spatial dependence and geographically weighted regression (GWR) and multiscale GWR (MGWR) models to locally examine spatial non-stationarity. The results suggested that even though incorporating spatial autocorrelation could significantly improve the performance of the global ordinary least square model, these models still represent a significantly poor performance compared to the local models. Moreover, MGWR could explain the highest variations (adj. R^2: 68.1%) with the lowest AICc compared to the others. Mapping the effects of significant explanatory variables (i.e., income inequality, median household income, the proportion of black females, and the proportion of nurse practitioners) on spatial variability of COVID-19 incidence rates using MGWR could provide useful insights to policymakers for targeted interventions.

* Corresponding author at: Department of Public Health & Prevention Science, Malicky Center, #328C, Baldwin Wallace University, 275 Eastland Road, Berea, OH 44017, USA.
E-mail addresses: amollalo@bw.edu (A. Mollalo), behzad@ucsb.edu (B. Vahedi), krivera19@bw.edu (K.M. Rivera).

https://doi.org/10.1016/j.scitotenv.2020.138884

1. Introduction

Coronavirus disease (COVID-19) caused by the SARS-CoV-2 virus, is a global health concern due to the rapid spread of the disease (WHO, 2020a). As of April 12, 2020, >105,000 deaths and nearly 1,700,000 incident cases have been globally confirmed (WHO, 2020b) and these figures are progressively increasing every day. The United Nations has described the disease as a *social, human, and economic crisis* (United Nations, 2020). The socioeconomic impacts and disease burden are especially evident in developing countries; however, the disease morbidity also impacts developed countries (United Nations, 2020). It is predicted that the annual global gross domestic product will decline by 24%, meaning that it is projected to decline by 2% each month (Congressional Research Service, 2020). The predictions also estimate a 13% to 32% decline in global trade (Congressional Research Service, 2020).

According to the World Health Organization (WHO, 2020c), COVID-19 was initially discovered in Wuhan, China, towards the end of 2019 before an outbreak of the disease was declared in January 2020. On March 11, 2020, the WHO officially declared the COVID-19 pandemic (WHO, 2020c). Shortly after, Iran and a few European countries, most notably Italy, experienced a significant increase in the number of cases and deaths (WHO, 2020c). In the United States, the first COVID-19 case was confirmed on January 19, 2020, in Washington State (Holshue et al., 2020). Thereafter, multiple states experienced an increased number of COVID-19 cases; New York State became one of the epicenters of the disease spread (Center for Infectious Disease Research and Policy, 2020). On March 17, 2020, all fifty states across the United States had confirmed cases of COVID-19 (Abir et al., 2020). On March 26, 2020, the United States became the leading country in the number of cases worldwide, replacing Italy that was previously in the lead of COVID-19 cases (Center for Infectious Disease Research and Policy, 2020). As of April 12, 2020, >20,000 deaths and >500,000 cases have been confirmed in the United States (The COVID Tracking Project, 2020).

Recent studies across the world have shown that multiple factors such as air pollution (Wu et al., 2020), smoking (Taghizadeh-Hesary and Akbari, 2020), and environmental conditions (Wang et al., 2020) may contribute to the severity and rate of spread pertaining to COVID-19. For example, Wu et al. (2020) showed that long-term air pollution exposure could potentially exacerbate the health outcomes of COVID-19 cases. Their findings also suggest that those with pre-existing conditions and air pollution exposure may suffer from higher mortality risk. In Iran, Taghizadeh-Hesary and Akbari (2020) suggest that smoking can negatively affect the health outcomes of COVID-19 patients due to potential decreased immune response. In China, Wang et al. (2020) indicated that environmental conditions such as humidity and temperature could influence the transmission of COVID-19 when compared to other respiratory viruses, suggesting a decline in disease spread.

Geographic information system (GIS) is an essential tool to examine the spatial distribution of infectious diseases (Mollalo et al., 2018, 2019), which can aid in the process of combating a pandemic and improving the quality of care (Lovett et al., 2014). GIS has become a vital tool in analyzing and visualizing the spread of COVID-19. For instance, Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) currently utilizes a GIS dashboard that provides live data of the worldwide spatial distribution of COVID-19, including the total number of confirmed cases, mortalities, and recovered patients (JHU CSSE, 2020). This nearly real-time database is readily accessible to the public, where they can keep track of the disease spread over time. The worldwide GIS map also accounts for the number of confirmed cases classified by country (JHU CSSE, 2020).

A limited number of GIS-based studies have been published since the initial outbreak of COVID-19. Boulos and Geraghty (2020), presented how various GIS applications and dashboards such as JHU CSSE, WHO dashboard, HealthMap, WorldPop, and EpiRisk are able to provide a clear representation of the COVID-19 spread. Lakhani (2020) utilized GIS mapping to identify COVID-19 health care priority locations pertaining to vulnerable populations, including elderly, palliative, and disabled patients in Melbourne, Australia. The findings suggest potential improvements in quality of care in the midst of the pandemic. Gibson and Rush (2020), utilized GIS technology to outline dwelling boundaries to detect the probability of COVID-19 spread in Cape Town, South Africa. Their results suggest that COVID-19 spread can be reduced through social distancing measures as supported by their buffer analysis and cluster identifications.

Spatial models are critical tools to statistically investigate the geographic relationship between several explanatory variables and disease outbreak (Mollalo et al., 2015; Mollalo and Khodabandehloo, 2016), such as COVID-19. In this study, we examine a few regressive and autoregressive spatial models to determine how well they can explain variations of COVID-19 in the continental United States based on several environmental, topographic, socioeconomic, behavioral, and demographic factors as explanatory variables. To our best knowledge, this paper provides the first attempt to use local geographic modeling of COVID-19 distribution across the United States and can provide useful insights for policymakers for targeted interventions.

2. Materials and methods

2.1. Data collection and preparation

The Centers for Disease Control and Prevention (CDC) continue to monitor state and county-level data of novel Coronavirus disease-daily and across the United States. For this study, the county-level counts of COVID-19 cases across the continental United States from January 22, 2020, to April 9, 2020, were retrieved from USAFacts (usafacts.org). Further, crude incidence rates were computed for the counties and joined to the administrative boundary shapefile of counties obtained from the TIGER/ Line database (www.census.gov) using ArcGIS Desktop 10.7.

A variety of 35 socioeconomic, behavioral, environmental, topographic, and demographic factors were compiled and considered as explanatory variables. Table 1 provides variable names together with their descriptions and the source of data. All variables were collected or prepared at the county-level and joined to the corresponding counties in ArcGIS environment.

To examine the relationship between the potential explanatory variables and the dependent variable (COVID-19 incidence rate), we used five different models. The models include three global models: ordinary least squares (OLS), spatial lag model (SLM), spatial error model (SEM), and two local models: geographically weighted regression (GWR), and multiscale GWR (MGWR).

2.2. Global models

2.2.1. Ordinary least squares (OLS)

The OLS is a regression method that investigates the relationships between a set of explanatory or independent variables and a dependent variable and has the general form of (Ward and Gleditsch, 2018):

$$\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{x_i}\boldsymbol{\beta} + \boldsymbol{\varepsilon_i} \tag{1}$$

where at county i, y_i is the COVID-19 incidence rates, β_0 is the intercept, x_i is the vector of selected explanatory variables, β is the vector of regression coefficients, and ε_i is a random error term. OLS optimizes regression coefficients (β) by minimizing the sum of squared prediction errors (Anselin and Arribas-Bel, 2013). OLS uses two major, implicit assumptions: that the observations are independent and constant across the study area and that the error terms are not correlated (Anselin and Arribas-Bel, 2013; Oshan et al., 2020).

OLS assumes that the observations at the county-level are independent of each other and does not consider spatial dependence. In reality,

Table 1
Explanatory variables used in this study together with definitions and sources.

Theme	Variable Name	Description	Source
Socioeconomic	(1) Median household income (2) Income inequality (3) Uninsured (4) Unemployment rate (5) Food Insecurity (6) Fair or poor health	(2) The ratio of household income at the 80th percentile to income at the 20th percentile (2018) (3) Percentage of population under age 65 without health insurance (2018) (4) Number of people ages 16+ unemployed and looking for work (2018) (5) Food Environment Index (2018) (6) Percentage of adults that report fair or poor health (2018)	(1–2) Small Area Income and Poverty Estimates, American Community Survey, five-year Estimates (3) Small Area Health Insurance Estimates (4) Bureau of Labor Statistics (5) Map the Meal Gap (6) Behavioral Risk Factor Surveillance System
Behavioral	Adult smoking	Percentage of adults that reported currently smoking (2018)	Behavioral Risk Factor Surveillance System (BRFSS)
Environmental	(1) Road density (2) Particulate matter (PM) 2.5 (3) Air quality index (AQI) (4) Temperature (5) Precipitation	(1) The total length of primary and secondary roads for each county calculated/area of the corresponding county (2) Daily minimum, maximum and average (3) Minimum, maximum and average AQI; (4) Minimum, maximum and average temperature; (5) Total precipitation	(1) US Census Bureau TIGER/Line (2–3) US Environmental Protection Agency (EPA) (4-5) National Oceanic and Atmospheric Administration (NOAA)
Topographic	(1) Minimum, maximum, and average (2) Maximum slope	(1) Digital elevation model of the United States (1 km spatial resolution)	United States Geological Survey (USGS)
Demographic	(1) Percent of 65 years and over (2) Percent of Asian (3) Percent of Hispanic (4) The proportion of African American (5) Percent of black males and females; (6) Percent of white males and females (7) Net International migration rate (8) Total number of primary care physicians* (9) Total number of nurse practitioners* (10) Total number of physician assistants* (11) Total number of hospitals	*Assumed proportion to the fraction of state population living in the county	(1–7) US Census Bureau Population Estimates (2018) (8–10) Healthcare Capacity including Physicians, Nurse Practitioners, and Physician Assistants (2019) (11) Kaiser Family Foundation and AAMC

however, and in the case of COVID-19 spread, we know that variables are spatially correlated (as supported by the results of SEM and SLM later on). These interactions are omitted from OLS, and therefore OLS is a misspecified model in this case (Anselin and Arribas-Bel, 2013). Thus, we used SLM and SEM that are both variants of OLS (Anselin, 2003; Ward and Gleditsch, 2018) and both take spatial dependence into account, but model it differently.

2.2.2. Spatial lag model (SLM)

The SLM assumes dependency between the dependent variable and explanatory variables and incorporates spatial dependence into the regression model with a "spatially-lagged dependent variable" (Anselin, 2003; Ward and Gleditsch, 2018). SLM is denoted by:

$$\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{x_i\beta} + \boldsymbol{\rho W_i y_i} + \boldsymbol{\varepsilon_i} \tag{2}$$

where ρ is the spatial lag parameter (spatial autoregressive parameter), and W_i is a vector of spatial weights (a row of the spatial weights matrix). Eq. (2) is constructed by decomposing the error term in Eq. (1) (Ward and Gleditsch, 2018). The weight matrix (W) on the right-hand side of this equation specifies the neighbors at location i and, as such, relates the independent variable to the explanatory variables at that location (Anselin and Arribas-Bel, 2013). The presence of spatial lag suggests a potential diffusion process (Kostov, 2010).

2.2.3. Spatial error model (SEM)

The SEM assumes spatial dependence in the error term of OLS and decomposes the error term in Eq. (1) into two terms ($\lambda W_i \xi_i$ and ε_i below) (Anselin, 2003; Chen et al., 2016). The general form of this model is: (Ward and Gleditsch, 2018)

$$\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{x_i\beta} + \boldsymbol{\lambda W_i \xi_i} + \boldsymbol{\varepsilon_i} \tag{3}$$

where at county i, ξ_i indicates the spatial component of the error, λ indicates the level of correlation between these components, and ε_i is a spatially uncorrelated error term.

2.3. Local models

2.3.1. Geographically weighted regression (GWR)

Global regression models such as OLS, SEM, and SLM implicitly assume spatial stationarity in the relationships between explanatory variables and dependent variable(s), meaning that they assume these relationships do not vary over space (Brunsdon et al., 1996; Brunsdon et al., 1998). To relax this assumption and to allow for "parameters to vary spatially." Brunsdon et al. (1996) introduced GWR as an extension of general regression models and based on kernel-weighted regression. Instead of estimating global values for regression parameters, GWR allows these parameters to be derived for each location separately, and in doing so, it incorporates geographic context (Oshan et al., 2020). GWR is denoted by (Fotheringham and Oshan, 2016)

$$\boldsymbol{y_i} = \boldsymbol{\beta_{i0}} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} \boldsymbol{\beta_{ij} X_{ij}} + \boldsymbol{\varepsilon_i}, \boldsymbol{i} = \mathbf{1}, \mathbf{2}, \boldsymbol{\ldots}, \boldsymbol{n} \tag{4}$$

where at county i, y_i is the value for the COVID-19 incidence rate, β_{i0} is the intercept, β_{ij} is the jth regression parameter, X_{ij} is the value of the jth explanatory parameter, and ε_i is a random error term. Parameter

Table 2
Summary statistics of the OLS model on selected variables in modeling COVID-19 incidence rates, continental United States.

Variable	Coefficient	T-statistic	P-value	VIF
Intercept	0.0007	0.0397	0.968338	–
Income inequality	0.2021	9.9015	0.000000*	1.4657
Median household income	0.2449	12.2474	0.000000*	1.4066
% of nurse practitioner	0.1365	7.4003	0.000000*	1.1963
% of black females	0.1095	5.7726	0.000000*	1.2667

estimates for each explanatory variable and at each county in matrix form is given by (Fotheringham and Oshan, 2016):

$$\hat{\beta}(i) = \left(X'W(i)X\right)^{-1}X'W(i)\,y \tag{5}$$

where $\hat{\beta}$ is the vector of parameter estimates $(m \times 1)$, X is the matrix of the selected explanatory variables $(n \times m)$, $W(i)$ is the matrix of spatial weights $(n \times n)$, and y is the vector of observations of the dependent variable $(m \times 1)$ (Fotheringham and Oshan, 2016). $W(i)$ is a diagonal matrix that is constructed from the weights of each observation based on its distance from location i and is calibrated based on a locally weighted regression (Brunsdon et al., 1998; Fotheringham and Oshan, 2016). To calculate $W(i)$, a kernel function and a bandwidth should be specified. The most commonly used kernel functions are Gaussian, and bi-square and the bandwidth is usually determined based on (Euclidean) distance or the number of nearest neighbors. Note that selecting different bandwidth types would affect the type of neighborhood in which local weighting happens.

2.3.2. Multiscale GWR (MGWR)

Even though GWR can be a great improvement compared to global regression in the context of spatial processes, it still assumes that the scale of all of the involved relationships are constant over space and thus does not allow for analyzing these relationships at different scales (Fotheringham et al., 2017; Oshan et al., 2019). Whereas, in many cases, including COVID-19 spread, this assumption is not valid because different processes are involved with varying spatial scales.

MGWR is an extension of GWR that allows studying the relationships at varying spatial scales and achieves that by using varying bandwidth as opposed to a single, constant bandwidth for the entire study area (Fotheringham et al., 2017; Yu et al., 2019). MGWR can be formulated as (Fotheringham et al., 2017):

$$y_i = \sum_{j=0}^{m} \beta_{bwj} X_{ij} + \varepsilon_i, i = 1, 2, \ldots, n \tag{6}$$

where β_{bwj} is the bandwidth used for calibration of the jth relationship (Fotheringham et al., 2017), and the rest of the parameters are the same as Eq. (1). In practice, MGWR is usually treated as a generalized additive model (GAM), which as a result, allows it to be calibrated using backfitting algorithms (Fotheringham et al., 2017; Hastie and Tibshirani, 1986; Buja et al., 1989). By reformulating MGWR as a GAM, we have:

$$y_i = \sum_{j=0}^{m} f_{ij} + \varepsilon \tag{7}$$

where f_{ij} (replacing $\beta_{bwj}X_j$ in (3)) is the jth additive term and is a smoothing function applied to jth explanatory variable at county i (Fotheringham et al., 2017; Oshan et al., 2019). Calibrating the model will result in a set of bandwidth, one for each of the j explanatory variables. Differences in bandwidths represent differences in spatial scales, and by capturing the effect of scale in spatial processes, MGWR can more accurately capture spatial heterogeneity (Fotheringham et al., 2017; Oshan et al., 2019).

2.4. Models development

Due to the existence of a relatively large set of candidate variables, the stepwise forward procedure was applied to select a subset of variables by eliminating non-significant explanatory variables. Subsequently, Pearson's correlation analysis was applied to investigate the correlations between all pairs of selected variables. Variance inflation factor (VIF) was used to detect multi-collinearity, and therefore the most uncorrelated factors were selected as the input of the models. For comparison, OLS and all the following models were implemented with the same selected variables. All global models were run in GeoDa 1.14 software (geodacenter.github.io). The weight matrix was generated based on first-order Queens' contiguity. Local models were implemented in MGWR 2.2 (https://sgsup.asu.edu/sparc/mgwr). An (adaptive) bi-square kernel, which removes the effect of observations outside the neighborhood specified with the bandwidth and (minimized) corrected Akaike Information Criterion (AICc), was used to select optimal bandwidth (Oshan et al., 2020; Oshan et al., 2019). The adjusted R^2 and AICc were used to compare the performances of models in explaining COVID-19 incidence rates across the continental United States.

3. Results

After feature selection and correlation analysis (correlation coefficients <0.3), among the 35 collected candidate variables, only four variables were selected to be included in the final models. These variables are income inequality, median household income, the percentage of nurse practitioners, and the percentage of the black female population (to the total female population) at the county-level (Table 2). As seen in Table 2, in the OLS model, the selected variables have relatively low multi-collinearity since the VIFs for all of them are below the threshold of 5 (all VIFs <1.5) (O'Brien, 2007) and were positively associated with COVID-19 incidence rates ($P < 0.001$). Although the global OLS model presented a very low adjusted R^2, it provided a baseline for subsequent global and local models. Low adjusted R^2 implies that almost 87.3% of the COVID-19 incidence rates across the continental United States are caused by unknown variables to the model and likely due to the local variations which were not captured by the OLS model.

According to Table 3, by incorporating spatial dependence, SLM and SEM improve the performance of OLS in modeling the COVID-19

Table 3
Summary statistics of SLM and SEM in modeling COVID-19 incidence rates, continental United States.

Variable	Coefficient		Std. error		Z-score		P-value	
	SLM	SEM	SLM	SEM	SLM	SEM	SLM	SEM
Intercept	−0.002	−0.003	0.016	0.027	−0.134	−0.098	0.893	0.922
Income inequality	0.172	0.189	0.019	0.021	8.98	9.158	0.000	0.000
Median household income	0.183	0.237	0.019	0.023	9.58	10.396	0.000	0.000
% of nurse practitioner	0.078	0.066	0.017	0.019	4.54	3.446	0.000	0.001
% of black females	0.064	0.123	0.018	0.0251	3.57	4.905	0.000	0.000
Rho	0.0402	–	0.024	–	16.99	–	0.000	–
Lambda	–	0.415	–	0.024	–	17.099	–	0.000

Table 4
Measures of goodness-of-fit for OLS, SEM, SLM, GWR, and MGWR in modeling COVID-19 incidence rate, continental United States.

Criterion	OLS	SEM	SLM	GWR	MGWR
Adj. R^2	0.127	0.238	0.242	0.674	0.681
AICc	8304.98	8063.52	8045.70	6134.19	`5796.53

incidence rate in the United States. Both autoregressive lag coefficients (i.e., ρ and λ in Eqs. (2) and (3), respectively) were found strongly significant ($P < 0.000$). However, spatial lag achieved a lower standard error of estimated parameters. Although both SEM and SLM significantly outperformed OLS, they still showed relatively poor performances in modeling the COVID-19 incidence rates in the United States. As mentioned before, this could be due to the neglected scale of spatial processes involved in modeling the disease incidence rate (Table 4).

To test potential local spatial differences, (M)GWR were employed. According to Tables 3 and 4, the value of adjusted R^2 significantly increased from 24.2% in the SLM (the most accurate general model in this study) to 67.4% in the GWR model. Moreover, the AICc dropped from 8045.70 to 6134.19. Among the employed models, the MGWR model showed the lowest AICc value (AICc: 5796.53), indicating the most parsimonious model. Moreover, MGWR obtained the highest adjusted R^2 (0.681), suggesting that the model could explain 68.1% of the total variations of COVID-19 incidence rates. This measure of goodness-of-fit was slightly lower for regular GWR (Adj. R^2: 0.674), with higher AICc compared to MGWR (AICc: 6134.19).

Figs. 1 and 2 show the results of mapping coefficients of GWR and MGWR for the selected variables. As seen in Fig. 1, income inequality demonstrated almost similar patterns in describing the geographic distribution of COVID-19 incidence rates at the county-level in both GWR and MGWR. Income inequality was an influential factor in explaining disease incidence rates across counties in the tri-state area (i.e., New York, Connecticut, and New Jersey states), Massachusetts, and in parts of the Western United States, particularly in Nevada, Idaho, and Utah.

On the contrary, both models represented poor performances at counties in the Southern United States, particularly in Arizona, Texas, and the New Mexico States, and also in most of the Northern Great Plains, particularly in North Dakota, South Dakota, and the Montana States. Median household income also revealed almost similar patterns to income inequality in both GWR and MGWR models.

In both GWR and MGWR, the percentage of nurse practitioners was a substantial factor in describing the geographic distribution of COVID-19 incidence rates in a number of counties in Louisiana, southern Mississippi, and a few counties in The Central United States and Midwest (Fig. 2). However, the impact of the percentage of black females on COVID-19 incidence rates was inconsistent between GWR and MGWR models.

Fig. 3 illustrates the spatial distributions of local R^2 values in both GWR and MGWR models. In MGWR, several counties in southern Florida, southern Mississippi, eastern Wisconsin, and western California had very high local R^2, indicating a decent prediction of the model in these areas. On the contrary, the local R^2 values were low in most of the counties in Central and Southern United States, indicating the poor performance of the model across these counties. Although there is a clear consistency between the local goodness-of-fit of both GWR and MGWR, it is evident that MGWR was more conservative than GWR.

4. Discussion

In this GIS-based research, we compiled 35 variables that could potentially explain the spatial pattern witnessed in the COVID-19 incidence rate at the county-level across the continental United States. These variables were grouped into five different themes, namely socio-economic, environmental, behavioral, topographic, and demographic. An ensemble of these variables was used to model the geographic distribution of COVID-19 incidence using a family of spatial regression and autoregressive models. Based on our findings, a combination of four variables of median household income, income inequality, percentage of nurse practitioners, and percentage of black female population could explain a relatively high variability of the disease incidence in the

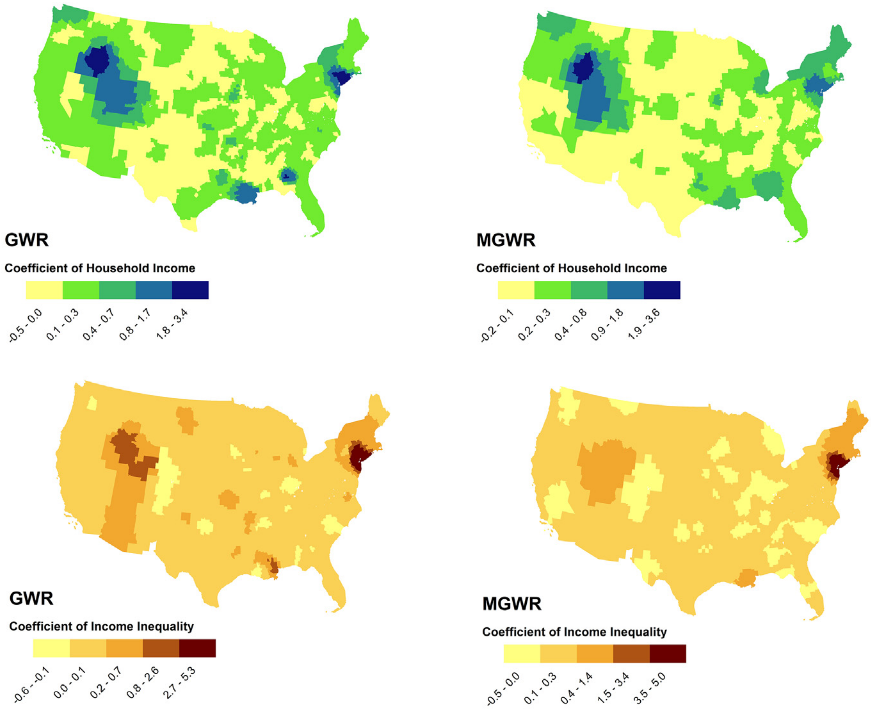


Fig. 1. The effects of median household income (above) and income inequality (below) in describing COVID-19 incidence rates using GWR (left) and MGWR (right) models, continental United States.

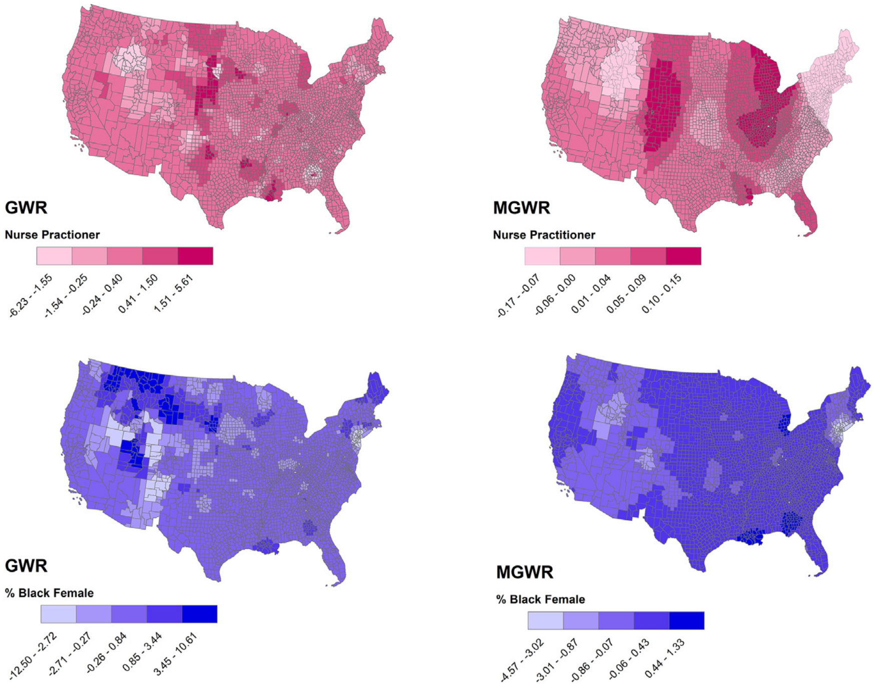


Fig. 2. The effects of % of nurse practitioners (above) and % of black females (below) in describing COVID-19 incidence rates using GWR (left) and MGWR (right), continental United States.

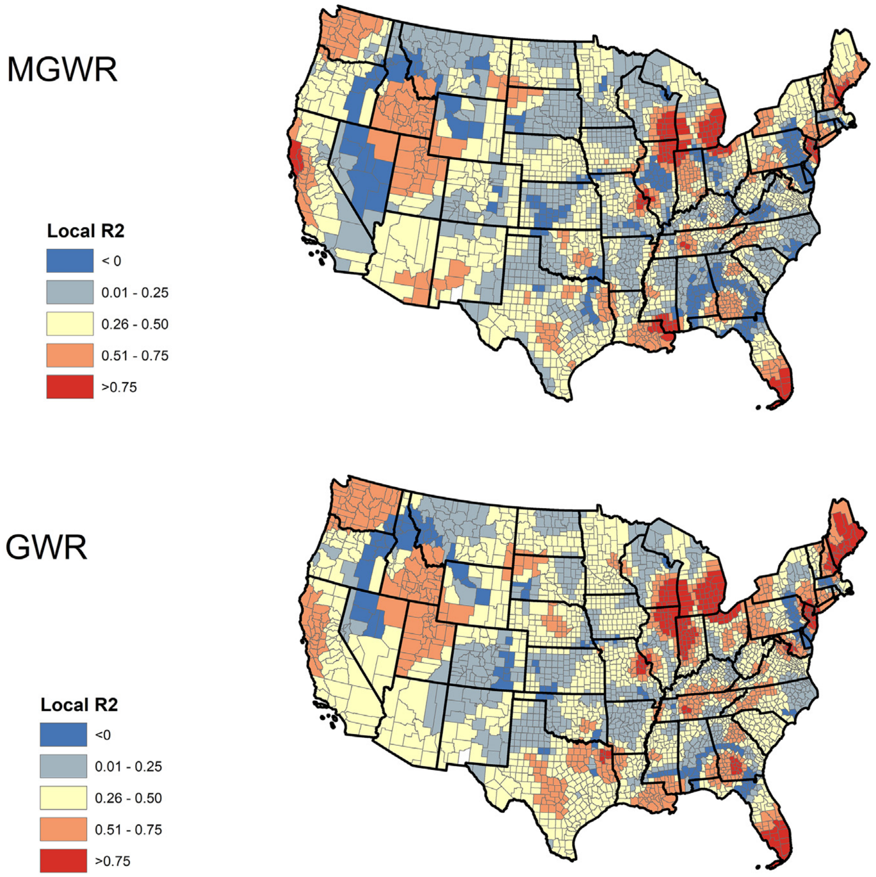


Fig. 3. Geographic distribution of local R^2 of GWR and MGWR models for COVID-19 incidence rate associated with income inequality, median household income, % of nurse practitioners, and % of black females across the continental United States.

continental United States. Continued monitoring of these factors can aid in understanding the dynamics of disease spread. Among the implemented models, MGWR was shown to better explain the spatial context of COVID-19 incidence rates. Through the use of variable bandwidths, MGWR allowed for modeling the effect of neighboring counties in variable neighborhood sizes and provided more flexibility in studying the extent of spatial processes.

At the time of writing this manuscript, the states of New York, New Jersey, Louisiana, Massachusetts, and Connecticut respectively have the highest incidence of COVID-19 per population in the United States. Findings of GWR and MGWR suggested a strong positive relationship of disease incidence with income inequality and median household income in these areas. Ahmed et al. (2020) allude to the socioeconomic disadvantages and inequalities that arise during pandemics; COVID-19 is not an exception. As the disease continues to spread, the world has witnessed substantial vulnerabilities in healthcare systems, a steep decline in economies, and an increase in unemployment rates. For example, in the United States, those who become unemployed are at risk of losing their health insurance coverage, which can directly contribute to the health and economic disparities that already exist in the country (Gangopadhyaya and Garrett, 2020) and as such this pandemic can cause a feedback loop.

Furthermore, our findings support the substantial impact of healthcare professionals, such as nurse practitioners, during the pandemic. For instance, a recent article emphasized the presence of a significant number of healthcare professionals within 55 years old or over who are working on the frontline (Buerhaus et al., 2020). Their results suggest the importance of continued training for younger health care professionals in the United States. Yet, nurse practitioners and physician assistants may be limited in their health care practice due to state law limitations across numerous states (Bayne et al., 2020). Although we did not find consistent data pertaining to demographics, Dowd et al. (2020) emphasizes the importance of considering population dynamics and demographic data to mitigate the approaches to combat the pandemic.

Based on our study, environmental factors did not demonstrate to be substantially influential when compared to COVID-19 incidence. However, in China, Ma et al. (2020) found a significant association with diurnal temperature range and COVID-19, particularly pertaining to daily mortality. Further studies may consider temperature anomalies to analyze the severity of COVID-19 across the continental United States. While we did not find smoking to be significantly influential, Brake et al. (2020) emphasize that smoking contributes to the vulnerability of combating COVID-19. Their findings also highlight that smoking may not be limited to traditional cigarettes; other smoking methods and devices are to be further investigated, including electronic cigarettes and waterpipe smoking (Brake et al., 2020).

One of the limitations of this study was data availability. Due to unprecedented efforts in the global research community to provide and share public data regarding different aspects of the COVID-19 pandemic, access to disease data is not difficult. However, to the best of our knowledge, the finest spatial granularity at which nationwide COVID-19 data in the United States is available is at the county-level. Therefore making inferences at the sub-county and individual levels may not necessarily produce accurate results. Another limitation was modeling different statewide shelter-in-place or lockdown policies (or lack thereof) and the level at which such policies were implemented and enforced. Different states have had variations in policies and approaches ranging from a relatively early shelter in place orders in states such as New York and California to no limitations in Arkansas, Nebraska, and South Dakota. Such policies and their implementations could result in extraordinary impacts on disease incidence rates. However, isolating or modeling such effects would be a challenging task that was out of the scope of this study. Moreover, though we did not include pre-existing conditions as explanatory variables, they should be incorporated in further studies. Recent articles have considered comorbidities such as diabetes (Gupta et al., 2020) and cardiovascular conditions (Zheng et al., 2020) as potential risk factors for COVID-19. These risk factors may be significantly influential in COVID-19 health outcomes. Further analysis supporting the mentioned variables may aid in improving the quality of care, policy development, and an overall improvement in combating the pandemic.

5. Conclusions

Inspired by Oshan et al. (2019), who applied MGWR to study the spatial context of obesogenic process in the state of Arizona, and presuming that a multiscale approach would better explain the spatial variability of COVID-19 rate across the United States, we applied and compared the performance of MGWR to four other global or local models. Our results confirmed and extended the findings of the mentioned study as MGWR achieved the highest goodness-of-fit with the most parsimonious model, among others. The spatial variability of MGWR in different counties can reflect different behavior of COVID-19 incidence rates in response to the selected explanatory variables. To the best of our knowledge, there is a lack of nationwide researches on geographic modeling of COVID-19; thus, this study can be regarded as a basis for future geographic modeling of the diseases.

CRediT authorship contribution statement

Abolfazl Mollalo: Conceptualization, Data curation, Formal analysis, Writing - review & editing.**Behzad Vahedi:**Conceptualization, Writing - review & editing.**Kiara M. Rivera:**Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We would like to thank anonymous reviewers for taking the time and effort to review the manuscript. Behzad Vahedi would like to thank Pouria Mistani and Samira Pakravan for their useful discussions and help. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Abir, M., Nelson, C., Chan, E.W., Al-Ibrahim, H., Cutter, C., Patel, K., Bogart, A., 2020. Critical Care Surge Response Strategies for the 2020 COVID-19 Outbreak in the United States. Retrieved from RAND Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RRA100/RRA164-1/RAND_RRA164-1.pdf.

Ahmed, F., Ahmed, N., Pissarides, C., Stiglitz, J., 2020. Why inequality could spread COVID-19. Lancet Public Health https://doi.org/10.1016/S2468-2667(20)30085-2 (Published online ahead of print).

Anselin, L., 2003. Spatial externalities, spatial multipliers and spatial econometrics. Int. Reg. Sci. Rev. 26 (2), 153–166.

Anselin, L., Arribas-Bel, D., 2013. Spatial fixed effects and spatial dependence in a single cross-section. Pap. Reg. Sci. 92 (1), 3–17.

Bayne, Ethan, Norris, Conor, Timmons, Edward, 2020. A primer on emergency occupational licensing reforms for combating COVID-19. SSRN Electron. J. https://doi.org/10.2139/ssrn.3562340.

Boulos, M. N. K., & Geraghty, E. M. (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics.

Brake, S.J., Barnsley, K., Lu, W., McAlinden, K.D., Eapen, M.S., Sohal, S.S., 2020. Smoking Upregulates angiotensin-converting Enzyme-2 receptor: a potential adhesion site for novel coronavirus SARS-CoV-2 (Covid-19). J. Clin. Med. 9 (3), 841.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr. Anal. 28 (4), 281–298.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician) 47 (3), 431–443.

Buerhaus, P.I., Auerbach, D.I., Staiger, D.O., 2020. Older clinicians and the surge in novel coronavirus disease 2019 (COVID-19). JAMA https://doi.org/10.1001/jama.2020.4978 (Published online March 30, 2020).

Buja, A., Hastie, T., Tibshirani, R., 1989. Linear smoothers and additive models. Ann. Stat. 453–510.

Center for Infectious Disease Research and Policy, 2020. US COVID-19 cases surge past 82,000, highest total in world. Retrieved from. https://www.cidrap.umn.edu/news-perspective/2020/03/us-covid-19-cases-surge-past-82000-highest-total-world.

Chen, Y., Chang, K.T., Han, F., Karacsonyi, D., Qian, Q., 2016. Investigating urbanization and its spatial determinants in the central districts of Guangzhou, China. Habitat International 51, 59–69.

Congessional Research Service, 2020. Global Econoic Effects of COVID-19. Retrieved from. https://fas.org/sgp/crs/row/R46270.pdf.

Dowd, J.B., Rotondi, V., Andriano, L., Brazel, D.M., Block, P., Ding, X., Mills, M.C., 2020. Demographic Science Aids in Understanding the Spread and Fatality Rates of COVID-19. https://doi.org/10.1101/2020.03.15.20036293.

Fotheringham, A.S., Oshan, T.M., 2016. Geographically weighted regression and multicollinearity: dispelling the myth. J. Geogr. Syst. 18 (4), 303–329.

Fotheringham, A.S., Yang, W., Kang, W., 2017. Multiscale geographically weighted regression (MGWR). Annals of the American Association of Geographers 107 (6), 1247–1265.

Gangopadhyaya, A., Garrett, B., 2020. Unemployment, Health Insurance, and the COVID-19 Recession. Urban Institute.

Gibson, L., Rush, D., 2020. Novel coronavirus in Cape Town informal settlements: feasibility of using informal dwelling outlines to identify high risk areas for COVID-19 transmission from a social distancing perspective. JMIR Public Health Surveill. 6 (2), e18844.

Gupta, R., Ghosh, A., Singh, A.K., Misra, A., 2020. Clinical considerations for patients with diabetes in times of COVID-19 epidemic. Diabetes & metabolic syndrome 14 (3), 211–212 Advance online publication. https://doi.org/10.1016/j.dsx.2020.03.002.

Hastie, T., Tibshirani, R., 1986. Generalized additive models. Stat. Sci. 1 (3), 297–310.

Holshue, M.L., DeBolt, C., Lindquist, S., Lofy, K.H., Wiesman, J., Bruce, H., ... Diaz, G., 2020. First case of 2019 novel coronavirus in the United States. New England Journal of Medicine 382, 929–936. https://doi.org/10.1056/NEJMoa2001191.

Johns Hopkins University Center for Systems Science and Engineering, 2020. COVID-19 Dashboard. Retrieved from. https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6.

Kostov, P., 2010. Model boosting for spatial weighting matrix selection in spatial lag models. Environment and Planning B: Planning and Design 37 (3), 533–549.

Lakhani, A., 2020. Which Melbourne metropolitan areas are vulnerable to COVID-19 based on age, disability and access to health services? Using spatial analysis to identify service gaps and inform delivery. J. Pain Symptom Manag. S0885-3924 (20), 30194–30199. https://doi.org/10.1016/j.jpainsymman.2020.03.041 (Published online ahead of print).

Lovett, D.A., Poots, A.J., Clements, J.T., Green, S.A., Samarasundera, E., Bell, D., 2014. Using geographical information systems and cartograms as a health service quality improvement tool. Spatial and Spatio-temporal Epidemiology 10, 67–74.

Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., ... Luo, B., 2020. Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Science of The Total Environment 724. https://doi.org/10.1016/j.scitotenv.2020.138226.

Mollalo, A., Khodabandehloo, E., 2016. Zoonotic cutaneous leishmaniasis in northeastern Iran: a GIS-based spatio-temporal multi-criteria decision-making approach. Epidemiology & Infection 144 (10), 2217–2229.

Mollalo, A., Alimohammadi, A., Shirzadi, M.R., Malek, M.R., 2015. Geographic information system-based analysis of the spatial and spatio-temporal distribution of zoonotic cutaneous leishmaniasis in Golestan Province, north-east of Iran. Zoonoses Public Health 62 (1), 18–28.

Mollalo, A., Sadeghian, A., Israel, G.D., Rashidi, P., Sofizadeh, A., Glass, G.E., 2018. Machine learning approaches in GIS-based ecological modeling of the sand fly Phlebotomus papatasi, a vector of zoonotic cutaneous leishmaniasis in Golestan province, Iran. Acta Trop. 188, 187–194.

Mollalo, A., Mao, L., Rashidi, P., Glass, G.E., 2019. A GIS-based artificial neural network model for spatial distribution of tuberculosis across the continental United States. Int. J. Environ. Res. Public Health 16 (1), 157.

O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. Qual. Quant. 41 (5), 673–690.

Oshan, T.M., Li, Z., Kang, W., Wolf, L.J., Fotheringham, A.S., 2019. Mgwr: a Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. ISPRS Int. J. Geo Inf. 8 (6), 269.

Oshan, T.M., Smith, J.P., Fotheringham, A.S., 2020. Targeting the spatial context of obesity determinants via multiscale geographically weighted regression. Int. J. Health Geogr. 19 (1), 1–17.

Taghizadeh-Hesary, F., Akbari, H., 2020. The powerful immune system against powerful COVID-19: a hypothesis. Preprints 2020, 2020040101. https://doi.org/10.20944/preprints202004.0101.v1.

The COVID Tracking Project, 2020. . Retrieved from. https://covidtracking.com/data.

United Nations, 2020. The Social Impact of COVID-19. Retrieved from. https://www.un.org/development/desa/dspd/2020/04/social-impact-of-covid-19/.

Wang, Jingyuan, Tang, Ke, Feng, Kai, Lv, Weifeng, 2020. High Temperature and High Humidity Reduce the Transmission of COVID-19.

Ward, M.D., Gleditsch, K.S., 2018. Spatial regression models. 155. Sage Publications.

World Health Organization (WHO), 2020a. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). Retrieved from. https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.

World Health Organization (WHO), 2020b. Coronavirus Disease 2019 (COVID-19) Situation Report – 83. Retrieved from. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200412-sitrep-83-covid-19.pdf?sfvrsn=697ce98d_4.

World Health Organization (WHO), 2020c. Rolling Updates on Coronavirus Disease (COVID-19). Retrieved from. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen.

Wu, X., Nethery, R.C., Sabath, B.M., Braun, D., Dominici, F., 2020. Exposure to Air Pollution and COVID-19 Mortality in the United States. https://doi.org/10.1101/2020.04.05.20054502.

Yu, H., Fotheringham, A.S., Li, Z., Oshan, T., Kang, W., Wolf, L.J., 2019. Inference in multiscale geographically weighted regression. Geogr. Anal. 52, 87–106.

Zheng, Y.Y., Ma, Y.T., Zhang, J.Y., Xie, X., 2020. COVID-19 and the cardiovascular system. Nat. Rev. Cardiol. 1–2.