Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental United States: A spatial nonlinear analysis

Yaowen Luo
Wuhan University
Jianguo Yan (jgyan@whu.edu.cn)
Wuhan University
Stephen McClure
Wuhan University

Research Article

Keywords: COVID-19 death rate; environment; socioeconomic; health; local nonlinear model; spatial variation

Posted Date: August 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-61369/v1

Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental United States: A spatial nonlinear analysis

Yaowen Luo[a,b]*， Jianguo Yan [b] and Stephen McClure[b]

[a] *Electronic Information School, Wuhan University, Wuhan, China*

[b] *State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China*

*Corresponding author email: luoyw@lreis.ac.cn

Address:

Electronic Information School Wuhan University
127 Luoyu Road
Wuhan, Hubei 430070

Abastract

The COVID-19 outbreak has become a global pandemic. Spatial variation in the environmental, health, socioeconomic, and demographic risk factors of COVID-19 death rate is not well understood. Global models and local linear models were used to estimate the impact of risk factors of the COVID-19, but these do not account for the nonlinear relationships between the risk factors and the COVID-19 death rate at various geographical locations. We proposed a local nonlinear nonparametric regression model named geographically weighted random forest (GW-RF) to estimate the nonlinear relationship between COVID-19 death rate and 47 risk factors derived from US Environmental Protection Agency, National Center for Environmental Information, Centers for Disease Control and the US census. The COVID-19 data were employed to a global regression model random forest (RF) and a local model GW-RF. The adjusted R^2 of the RF is 0.69. The adjusted R^2 of the proposed GW-RF is 0.78. The result of GW-RF showed that the risk factors (i.e. going to work by walking, airborne benzene concentration, householder with a mortgage, unemployment, airborne $PM_{2.5}$ concentration and percent of the black or African American) have a high correlation with the spatial distribution of the COVID-19 death rate and these key factors driven from the GW-RF were mapped, which could provide useful implications for controlling the spread of COVID-19 pandemic.

Keywords: COVID-19 death rate; environment; socioeconomic; health; local nonlinear model; spatial variation

1. Instruction

The 2019 novel coronavirus disease (COVID-19) caused by SARS-CoV-2, is a rapidly spreading infectious disease that mainly affects the respiratory system (Landi et al. 2020). Because the disease is highly contagious with rapid transmission between human (Huang et al. 2020), World Health Origination (WHO) declared on March 11 2020 that the COVID-19 outbreak has been a global pandemic (World Health Organization 2020b). As of July 6, 2020, a total of 11 520 953 COVID-19 confirmed cases and 532 633 deaths were recorded worldwide. The current epicentre of the COVID-19 is the United States with 2 982 928 confirmed cases and 132 569 deaths as if July 6, 2020. The economic impact from the COVID-19 crisis are unprecedented in United States with the substantial stock market shifting and unemployment rate reaching the peak (O'Connor et al. 2020). The health care system is also overwhelmed across the world, which are already operating at full capacity struggling to meet the demand for ventilators, intensive care beds and personal protective equipment.

Some researches about the COVID-19 have found that various factors including environment (Ahmadi et al. 2020; Bashir et al. 2020; Xu et al. 2020), socioeconomic (de León-Martínez et al. 2020; Zheng et al. 2020), demographic (Serge et al. 2020) and underlying disease (Dariya and Nagaraju 2020; Malik et al. 2020; Marhl et al. 2020; Ruthberg et al. 2020) may influence the transmission of COVID-19. Bashir et al. (2020) found that air pollution including PM_{10}, $PM_{2.5}$, SO_2, NO_2, and CO are significant risk factors to the COVID-19 epidemic. Tosepu et al. (2020) analysed the correlation between weather and the COVID-19, and found that the average temperature was highly correlated with the COVID-19. Virus carrier via public transportation played an important role in the transmission of COVID-19 (Zheng et al. 2020). Serge et al. (2020) found that males are about 60% more likely than females to suffer severe illness or death from the COVID-19 complications. Targher et al. (2020) found that patients with diabetes were at an approximately 4 times risk of having severe COVID-19.

With the increased availability of health care data online and the development of the spatial analysis techniques, multiple analysis by GIS tool(Guliyev 2020; Rosenkrantz et al. 2020) found that the distribution of COVID-19 cases (Desjardins et al. 2020; Lau et al. 2020; Shim et al. 2020) and its risk factors (Mollalo et al. 2020) exhibit patterns of spatial heterogeneity. A study by Lau et al. (2020) showed

that the number of flight routes was a highly relevant factor of the COVID-19 spread. Their study showed that regions in Asia, North America and Europe were at a serious risk of constant exposure to highly infected countries, while the exposure risk to the COVID-19 was relatively low in South America and Africa. Liu et al. (2020) employed a contact model to reconstruct the contact and air spread to simulate the outbreak of COVID-19 on the "Diamond Princess". They suggested rigorous prevention measure should be followed by the high-risk susceptible people. Mollalo et al. (2020) mapped the spatial variability of the relationships between COVID-19 incidence rate and income inequality, median household, the proportion of black females and proportion of nurse practitioners using multiscale geographically weighted regression (MGWR).

Many mathematical models have been employed to explore the risk factors of COVID-19. Typical global models such as Partial correlation coefficient (PCC) (Ahmadi et al. 2020), Ordinary least squares (OLS), Poisson regression model (Xu et al. 2020), Bayesian hierarchical model (Millett et al. 2020) and geographical local model such as geographically weighted regression model (GWR) (Imran et al. 2015; Mollalo et al. 2020) were used to model the correlations between COVID-19 data and other impacting factors. However, the global model assumes the relationship between risk factors do not vary over space and is inconsistent with the imbalanced distribution of COVID-19. Although Spatial Error Model (SEM) and Spatial Lag Model (SLM) do consider spatial factors, they focus more on the analysis of spatial correlation and do not analyze the spatial variation of the relationships between variables in different regions from the perspective of spatial heterogeneity (Ahmadi et al. 2020). The GWR (Brunsdon et al. 2010; Fotheringham et al. 2002; Lu et al. 2017) as a local regression model can obtain the linear relationship between variables in different locations. However, the GWR is constructed based on multiple linear regression models, thus it is not suitable to estimate the nonlinear relationships between independent and dependent variables, and local multicollinearity exists when dealing with correlated variables (Wheeler and Tiefelsdorf 2005). The real relationship between risk factors and COVID-19 is complex and is not always linear. In order to explore the spatial variation of the nonlinear relationship between multiple risk factors and COVID-19, it is necessary to deal with the nonlinear situation in a local regression model.

In this study, we proposed a local nonlinear nonparametric regression method, Geographically Weighted Random Forest (GW-RF), to evaluate the relationship between COVID-19 death rate and multiple risk factors including air pollution, climate, landcover, disaster, health status, commuting to

work, socioeconomic and demographic indicators at county level across the continental United States. This paper first tries to explore the variation in the nonlinear relationships between multiple risk factors and COVID-19 death rate in different locations by using the GW-RF. We expect this study can provide scientific evidence for implementing control and prevention measure in COVID-19.

2. Materials and methods

2.1 Data and preparation

The county level daily COVID-19 death cases data and population data of 3108 counties of the continental United State from Jan 22, 2020 to June 26, 2020 were downloaded from the website of USA FACTS (https://usafacts.org/). The death rate at county level was calculated based on the daily COVID-19 death cases and population data. We selected 47 indicators including atmosphere, climate, landcover, disaster, health status, commuting to work, socioeconomic and demographic factors as independent variables to evaluate their correlation with the COVID-19 death rate. The indicators we selected and their meanings and sources are presented in Table 1. The shapefile of the selected 3108 counties was downloaded from geographical program of United States Census Bureau (https://www.census.gov/programs-surveys/geography.html).

Table 1. Definitions of indicators and sources

Theme	**Indicators**	**Indicator meaning**	**Source**
Atmosphere	Airborne $PM_{2.5}$ concentration	Annual average ambient concentrations of $PM_{2.5}$ in micrograms per cubic meter	United States Environmental Protection Agency (https://www.epa.gov/) and Centers for Diseases Control and Prevention (https://www.cdc.gov/)
	Airborne benzene concentration	Annual average concentration of benzene estimates in microgram per cubic meter	
	Airborne formaldehyde concentration	Annual average air concentration of formaldehyde estimates in microgram per cubic meter	
	Airborne acetaldehyde concentration	Annual average air concentration of acetaldehyde estimates in microgram per cubic meter	

	Airborne carbon tetrachloride concentration	Annual average air concentration of carbon tetrachloride estimates in microgram per cubic meter	
Climate	Air temperature	Average Daily Max Air Temperature (F)	National Center for Environmental Information (https://www.ncei.noaa.gov/)
	Precipitation	Average Daily Precipitation (mm)	
	Sunlight exposure	Annual average sunlight exposure measured by solar irradiance (kJ/m^2)	Centers for Diseases Control and Prevention (https://www.cdc.gov/)
	UV radiation exposure	Annual average daily dose of UV irradiance (J/m^2)	
Landcover	Landcover with water	Percent of land covered by water	
	Land cover with forest	Percent of land covered by forest	
Disaster	Drought	Number of weeks of moderate drought or worse per year	
	Flood	Percentage of people within fema designated flood hazard area	
Health status	Disability	Percentage of population aged 5 years and over with a disability	
	Asthma	Percent of adults diagnosed with asthma	
	Obese	Percentage of adults aged 18 years and over who were obese	
	Overweight	Percentage of adults aged 18 years and over who were overweight	
	Cancer	Number of people with lung and brouchus cancer per 1000000 population	
Commuting to work	Go to work by private transportation	Percentage of workers 16 years and over who drove alone (car, truck, or van)	United States Census Bureau (https://www.census.gov/en.html)
	Go to work by public transportation	Percentage of workers 16 years and over who go to work by public transportation (excluding taxicab)	

	Go to work by walking	Percentage of workers 16 years and over who go to work by walking	
	Work at home	Percentage of workers 16 years and over who worked at home	
	Mean travel time to work	Mean travel time to work (minutes) of the workers 16 years and over	
Socioeconomic	Health insurance	Percentage of population without health insurance	
	Householder with a mortgage	Percentage of household with a mortgage	
	Poverty	Percentage of population whose income is below the poverty level	
	Service occupations	Percentage of employed population 16 years and over with service occupations	
	Unemployment	Percentage of population 16 years and over unemployed	
	Hospital	Number of hospitals	Centers for Diseases Control and Prevention (https://www.cdc.gov/)
	Hospital beds	Number of hospital beds per 10000 population	
	People living in group quarter	Percentage of population living in group quarter	United States Census Bureau (https://www.census.gov/en.html)
	People living near a park	Percentage of population living within a half mile of a park	
	Householder with no internet access	Percentage of households with no internet access	
	Median household income		
	Mean household retirement income		
	Mean household cash public assistance income		

	Mean household Supplemental Security Income
Demographic	Percent of males
	Median age
	Percent of people under 18 years
	Percent of people 65 years and over
	Percent of the white race
	Percent of the black or African American
	Percent of American Indian and Alaska Native
	Percent of Asian
	Percent of native Hawaiian and other Pacific islander
	Percent of Hispanic or Latino

Due to the units of these 47 indicators are different, the indicators should be normalized before regression. The method is as follows:

$$X_{ki} = \frac{X_{ki} - \bar{X}_k}{\sigma_k} (i \in 1,2,\cdots 2056; k \in 1,2,\cdots,28) \quad (1)$$

where X_{ki} represents the normalized value of the kth indicator in the i th county, X_{ki} represents the original value of the kth indicator in the i th county; $\bar{X}_k$ represents the average value of the kth indicator; σ_k represents the standard deviation of kth index. The COVID-19 death rate and 47 indicators were joined to the county level shapefile for futher processing.

2.2 Nonlinear nonparametric model

2.2.1 Random forest (RF)

We selected the Random forest (RF) machine learning method (Breiman 2001) because it is nonparametric, it can easily learn nonlinear relationships and interactions from data without explicitly modelling them. RF is an ensemble of multiple decision trees. The decision tree is a non-parametric model which does not have a fixed structure. The decision tree grows according to the complexity of the input data in the learning process. The RF works well for high-dimensional variables with a relatively small number of samples, and can access variable importance (Grömping 2009). The algorithm flow of the RF is as follows.

(1) The n data sets $D_1, D_2, \cdots, D_n$ are extracted by repeatedly using the bootstrap method to randomly extract the whole data set D; and the corresponding n decision trees $H_1, H_2, \cdots, H_n$ are generated.

(2) At each node of the decision tree, randomly select m ($m < k$) variables from all the k variables of the decision tree, and each node is split using the selected m variables by the optimal segmentation method determined by a segmentation criterion.

(3) The value of m remains unchanged while the forest grows. Each tree grows to its largest extent without pruning until it cannot be split.

Thus the correlation between the decision trees in the forest decreases through a random selection of variables at each node of the tree and the optimal split of each node is determined by the selected variables only, instead of all variables. Each tree can grow to its largest extent without pruning. Therefore, the algorithm can deal with excessive redundant features and avoid over fitting.

In the first step in constructing the RF, whether with or without replacement, approximately 36.8% of the data samples are not used to grow the tree; these samples are the Out-Of-Bag (OOB) for the tree. The accuracy of the RF model can be estimated from the OOB data as presented by Equation (1):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{\hat{y}}_i\right)^2 \quad (1)$$

where N is the number of samples which from OBB data, y_i is the actual value of the ith sample, $\bar{\hat{y}}_i$ is the average prediction for the ith sample from all trees.

The overall sum of squares (SST) and coefficient of determination (R^2) are respectively defined in Equations (2) and (3):

$$\mathrm{SST} = \sum_{i=1}^{N}(y_i - \bar{y})^2 \quad (2)$$

$$R^2 = 1 - N\frac{MSE}{SST} \quad (3)$$

where $R^2 \in (0,1)$. The closer of the value of R^2 to 1, the better the regression performance of the GW-RF will be.

Variable importance can sort the independent (predictor) variables according to their degree of correlation to the dependent (response) variable. There are two popular methods to measure the variable importance in the RF, which are average impurity reduction (Gini importance) and Mean Square Error (MSE) reduction. Because the result of variable importance by impurity reduction is biased (Strobl et al. 2007), many researchers have verified and suggested choosing the MSE reduction method when permuting the variables (Ishwaran 2007; Strobl et al. 2008). The MSE reduction method uses the MSE value of the Out-Of-Bag (OOB) data to evaluate the variable importance (Cai et al. 2018). It is determined as follows:

(1) Calculate the MSE of the OBB data for each tree. For tree t, the MSE of OOB data is calculated by Equation (4):

$$MSE_t = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(y_i - \hat{y}_{i,t}\right)^2 \quad (4)$$

where N_t is the number of samples from OBB data in the tree t, $\hat{y}_{i,t}$ is the prediction for the ith sample of the tree t.

(2) Randomly replace the target variable j, and then the new value of the MSE of tree t is calculated by Equation (5):

$$MSE_t(j) = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(y_i - \hat{y}_{i,t}(j)\right)^2 \quad (5)$$

where $\hat{y}_{i,t}(j)$ is the prediction for the ith sample of the new tree t when randomly replace the target variable j.

(3) Calculate the difference between MSE_t and $MSE_t(j)$, and the MSE reduction is the variable importance for variable j of tree t. The MSE reduction of variable j of the whole forest is obtained as the average over MSE reduction of all n trees. The variable importance of variable j is expressed as in Equation (6):

$$\mathrm{VI}(j) = \mathrm{MSE}(j) = \frac{1}{n}\sum_{t=1}^{n}(MSE_t - MSE_t(j)) \quad (6)$$

2.2.2 Geographically weighted random forest (GW-RF)

In this section, a local nonlinear machine learning method, denoted as GW-RF, is proposed. The GW-RF is designed by integrating spatial weight matrix (SWM) and RF into a local regression analysis framework. The GW-RF inherits the merits of the RF, making the RF from being applicable from a global system to a local system. Thus, it can handle high-dimensional variables with nonlinear relationships and multicollinearity. The variable importance for each spatial unit can be obtained from the GW-RF. The process of constructing the GW-RF model is designed as follows:

(1) The SWM for each spatial unit of the study area should first be made according to the specified spatial weight rule. The SWM for the whole study area with p spatial units can be expressed as in Equation (7):

$$W = \begin{bmatrix} W(1) \\ W(2) \\ \vdots \\ W(i) \\ \vdots \\ W(p) \end{bmatrix} = \begin{bmatrix} w_{11}\, w_{12} & \cdots & w_{1p} \\ w_{21}\, w_{22} & \cdots & w_{2p} \\ \vdots\ \vdots & \vdots & \vdots \\ w_{i1}\, w_{i2} & \cdots & w_{ip} \\ \vdots\ \vdots & \vdots & \vdots \\ w_{p1}\, w_{p2} & \cdots & w_{pp} \end{bmatrix}, \quad i \in (1,2,\cdots,p) \quad (7)$$

As the local random forest of an individual unit need to consider the unit itself, the value of w_{ii} is set to 1 ($w_{ii} = 1$). According to the spatial weight rule, for spatial unit i, if sample j ($j \in (1,2,\cdots,p) \wedge i \neq j$) is a "neighbour" of unit i, the value of spatial weight between them is set to 1, that is, $w_{ij} = 1$. While spatial unit j is far away from spatial unit i, not a "neighbour" of spatial unit i, $w_{ij} = 0$.

(2) Select all the "neighbours" of each spatial unit according to the spatial weight matrix. For unit i, the "neighbors" of it can be selected from the special weight matrix W where $w_{ij} \neq 0, (j \in (1,2,\cdots,p) \wedge i \neq j)$.

(3) The spatial unit i and its "neighbours" are as the inputs to construct a local RF for unit i (RF (i)). By executing RF (i), the variable importance for spatial unit i can be computed.

(4) Repeat steps (2) and (3) to construct a local RF for each spatial unit in the study area and estimate the local variable importance for each spatial unit.

The nonlinear nonparametric models (RF, GW-RF) does not need to consider multicollinearity, and can analyse all independent variables without screening. R software (version 3.5.3, http://cran.r-project.org) was used to perform the regression analysis.

3. Results

All 47 indicators were employed to the nonlinear nonparametric models (RF, GW-RF). The adjusted fitting coefficient (R^2) of the RF was 0.69, while the adjusted R^2 of the GW-RF was 0.78, indicating that the regression result of the GW-RF was more accurate than that of the RF. The variable importance of a independent variable represents the correlation between the independent variable and the dependent variable and the higher value of the variable importance is, the stronger the correlation will be. The variable importance of 47 independent variables in modelling COVID-19 death rate using the RF is showed in Figure 1. The risk factors referring to socioeconomic are most correlated with COVID-19 death rarte, followed by risk factors referring to demographic, commuting to work, atmosphere, health status, landcover, disaster and climate. The variables including householder with a mortgage, going to work by walking, land cover with forest, hospital beds, overweight, percent of Hispanic or Latino, people living in group quarter and Airborne benzene concentration have a high correlation with the COVID-19 death rate.

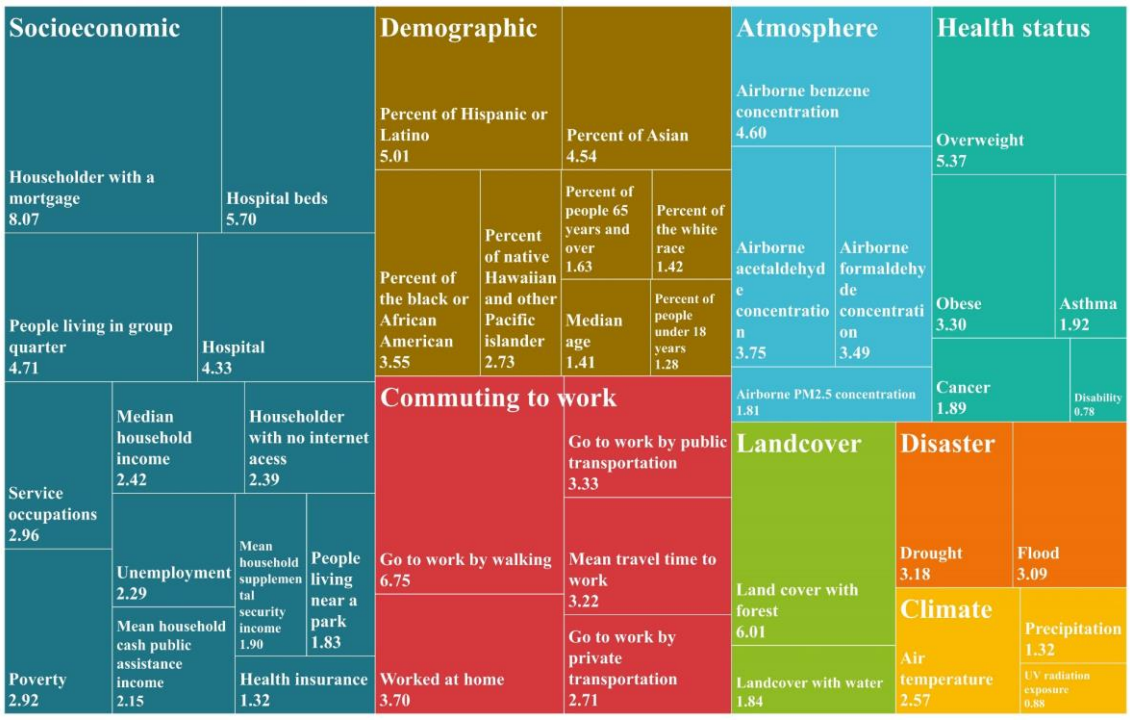

Figure 1. The variable importance of the independent variables of the RF model in modelling COVID-19 death rate

We used the local R^2 to estimate the performance of the GW-RF. Table 2 describes the statistic of local R^2 of the GW-RF. The average value of local R^2 was 0.59. The value of local R^2 was higher than 0.4 in 89.4% of the counties and higher than 0.6 in 50.5% of the counties. This shows that the GW-RF can accurately evaluate the correlation between the risk factors and the COVID-19 death rate in most of the study areas.

Table 2. The statistic of local R^2 of the GW-RF in modelling COVID-19 death rate, we calculated the average value of local R^2 and the percentage of counties in five local R^2 range (≤0.2, (0.2, 04], (0.4, 06], (0.6, 08], >0.8)

The value of local R^2	GW-RF
Average value	0.59
≤0.2	1.1%
(0.2, 04]	9.5%
(0.4, 06]	38.9%
(0.6, 08]	44.8%
>0.8	5.7%

Figure 2 shows the distribution of the local R^2 of the GW-RF across the study area. As can be seen from Figure 2, the distribution of local R^2 was imbalanced in the whole study area. The local R^2 value of the GW-RF was high in most of the counties across the whole continental United States, indicating that the GW-RF worked well in the prediction of the local COVID-19 death rate in most regions across the study area, especially in Nevada, Arizona, Washington and some counties in the East-central region.

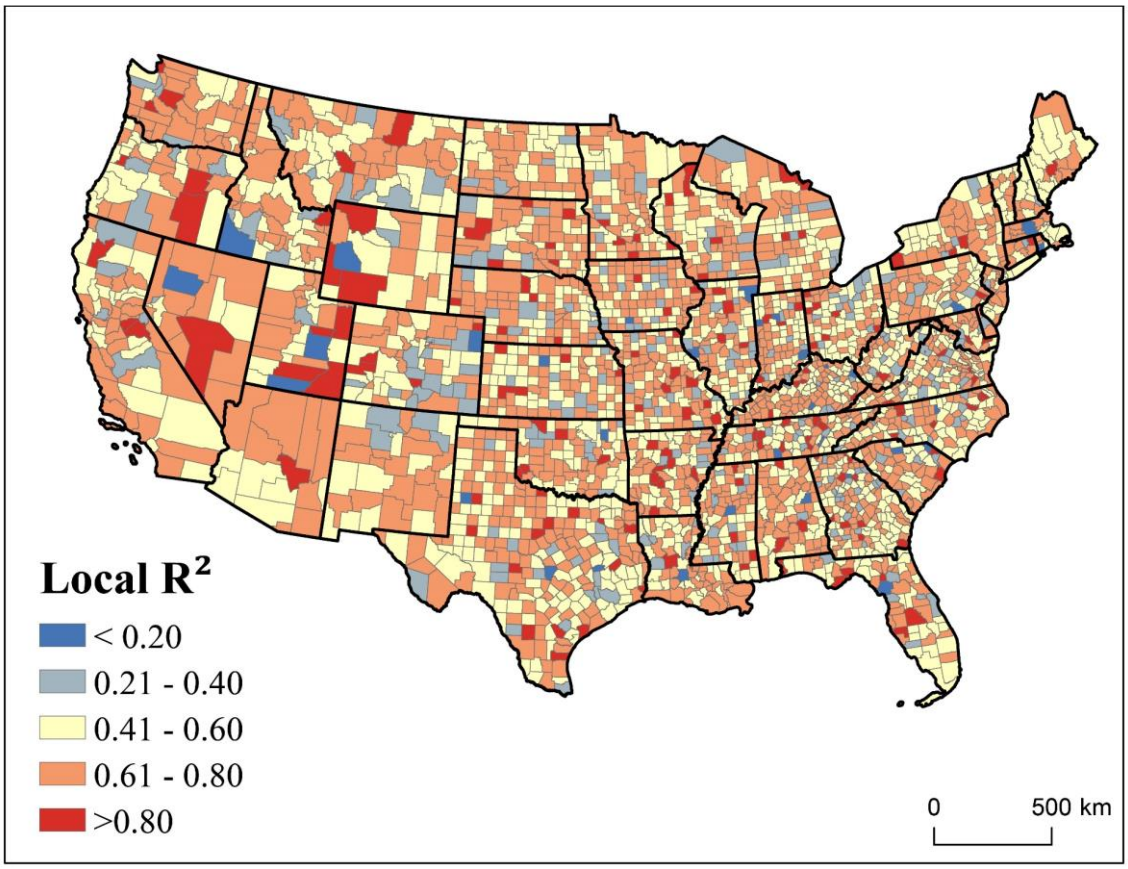

Figure 2. The distribution of local R^2 of the GW-RF

We computed the average local effect of each independent variable on COVID-19 death rate in the GW-RF model (see Figure 3). The effect of going to work by walking had the highest correlation with the COVID-19 death rate, flowed by airborne benzene concentration, householder with a mortgage, unemployment, airborne $PM_{2.5}$ concentration and percent of the black or African American.

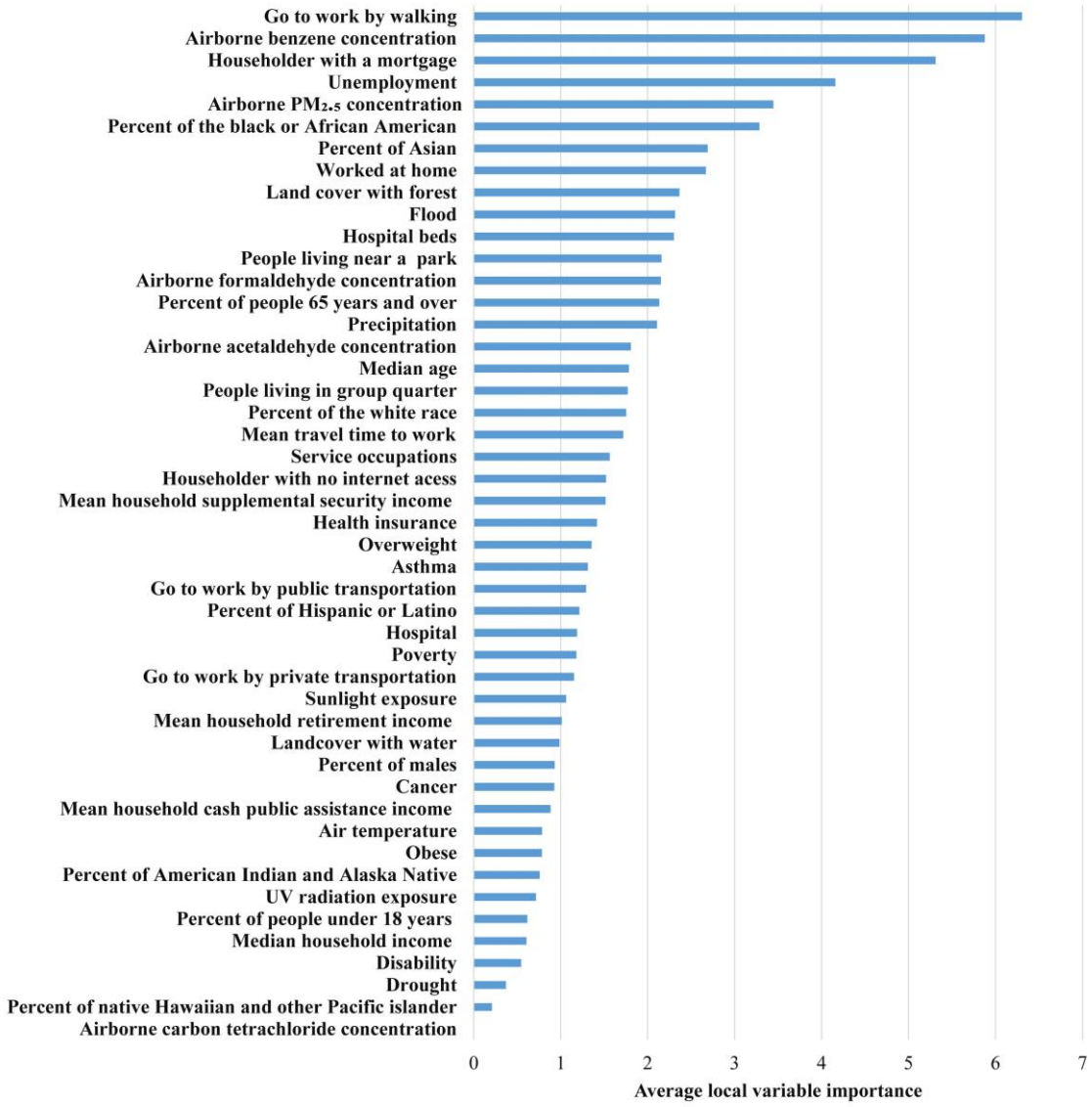

Figure 3. The average local variable importance of 47 potential risk factors on COVID-19 death rate in the GW-RF model

The proportion of counties with local primary risk factor (the risk factor with the highest value of local variable importance) at county level in the GW-RF was calculated (see Table 3). Going to work by walking was the most influential risk factor in 35% of the counties. The airborne benzene concentration was the leading risk factor in 24% of the counties. 13% percent of counties were most affected by householder with a mortgage and 12% percent of counties were most affected by unemployment. Figure 4, Figure 5 and Figure 6 provide a detailed spatial distribution of the local variable importance of first six factors with the highest value of average variable importance on the COVID-19 death rate using the GW-RF.

239 Table 3. The proportion of counties with local primary risk factor (the risk fator with the highest value of
240 local variable importance) on COVID-19 death rate at county level in the GW-RF.

Local primary risk factor	Proportion of counties
Go to work by walking	35%
Airborne benzene concentration	25%
Householder with a mortgage	13%
Unemployment	12%
Other risk factors	16%

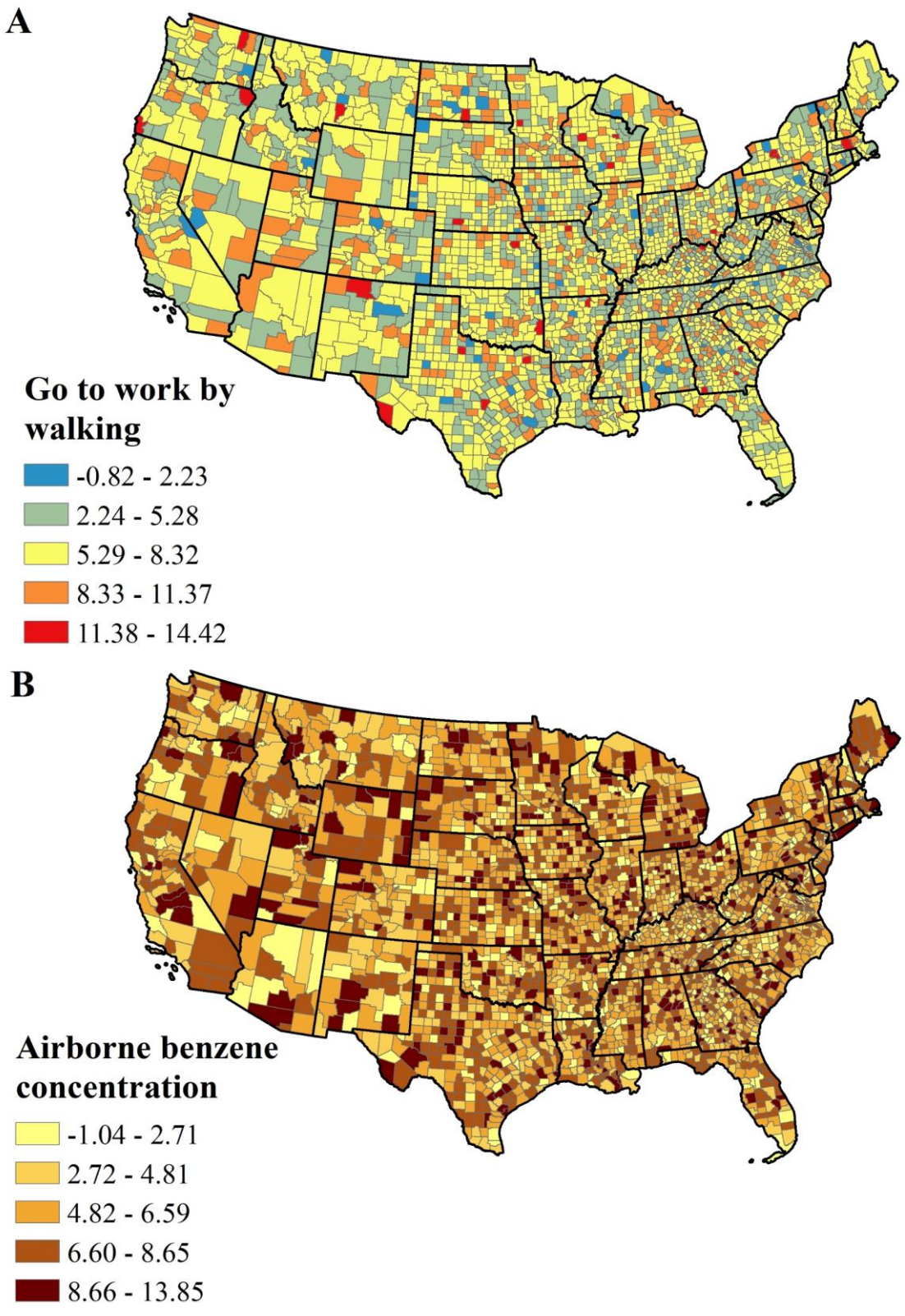

Figure 4. The spatial distribution of the local variable importance of (A) go to work by walking, (B) airborne benzene concentration on COVID-19 death rate in GW-RF model

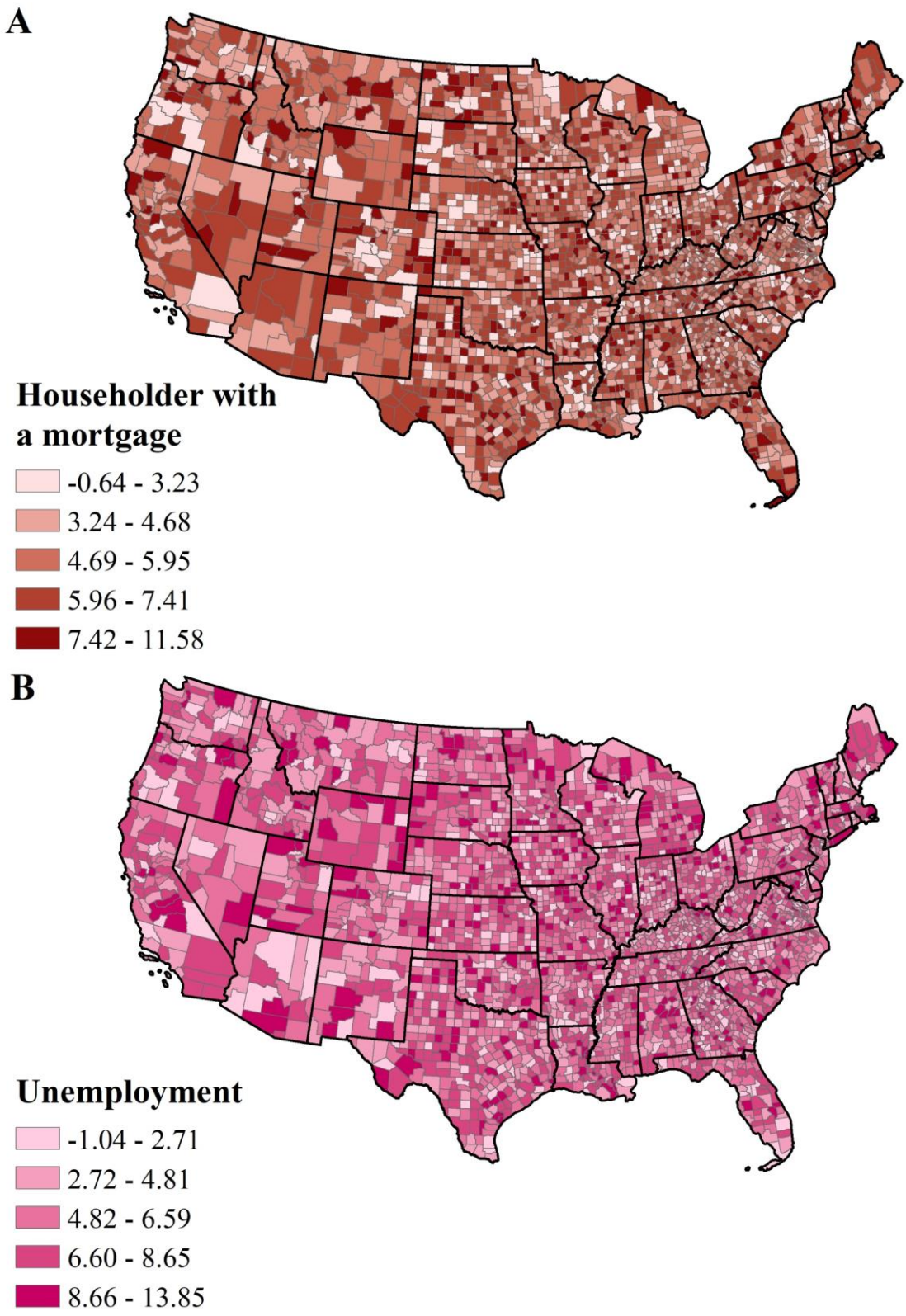

Figure 5 The spatial distribution of the local variable importance of (A) householder with a mortgage, (B) unemployment on COVID-19 death rate in GW-RF model

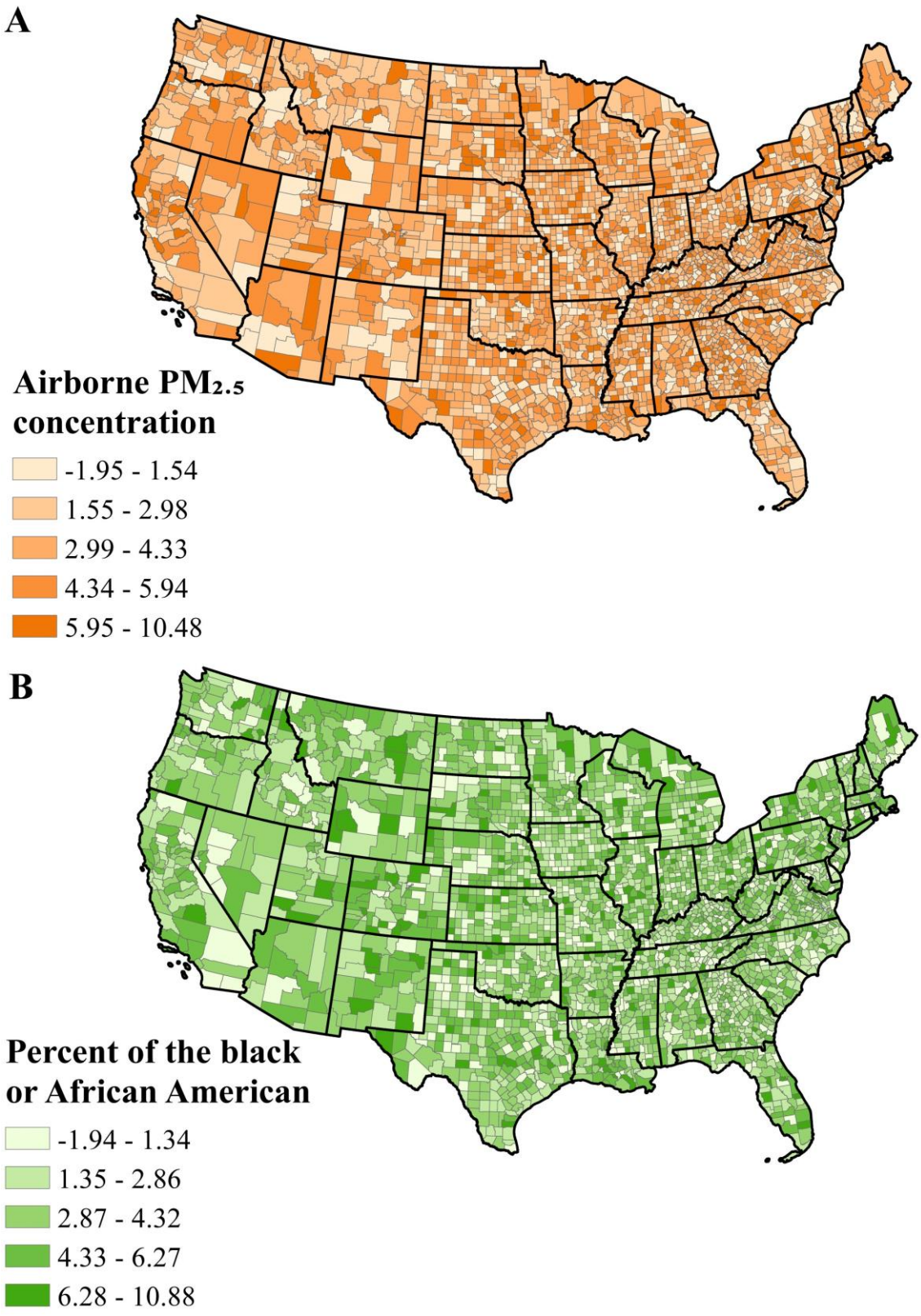

Figure 6 The spatial distribution of the local variable importance of (A) airborne $PM_{2.5}$ concentration, (B) percent of the black or African American on COVID-19 death rate in GW-RF model

From Figure 4, Figure 5 and Figure 6, the distribution of the variable importance of each variable on COVID-19 death rate in GW-RF model was imbalanced in different counties even the

counties in a same state. For example, in the southern part of Arizona, the COVID-19 death rate was mainly affected by the airborne benzene concentration and unemployment, and the northern part was mainly affected by going to work by walking and airborne $PM_{2.5}$ concentration. The regions obviously effected by going to work by walking were distributed in California, Arizona, the west of Utah, South Carolina and Massachusetts. The areas influenced by airborne benzene concentration were scattered throughout the study area. New Mexico, Florida, Texas, Missouri, the south of Nevada, the north of Arizona, Massachusetts and Connecticut were sensitive to householder with a mortgage. The regions obviously effected by airborne $PM_{2.5}$ concentration and percent of the black or African American are similar, mainly located in the north of Nevada, the north of Arizona, the southeast of Oregon, the east of Wyoming and the central part of the continental United States. In addition, the same area was affected by several risk factors. For example, airborne benzene concentration, householder with a mortgage, unemployment and the percent of black of African American were influential factors in the southeast of Arizona.

4. Discussion and conclusion

Identifying the risk factors that highly correlated with the transmission will provide guidance in containing the spread of the COVID-19 disease. In this study, we selected 47 potential risk factors from atmosphere, climate, land cover, disaster, health status, commuting to work, socioeconomic and demographic categories as independent variables to estimate their impact on the distribution of the COVID-19 death rate at county level across continental United States. Due to the imbalanced distribution of COVID-19 death rate and the complex relationship between the COVID-19 death rate and its risk factors, the linear models could not accurately identify the key risk factors in different locations. To solve this problem, we applied GW-RF, a local regression model capable of identifying nonlinear relationships between variables at various geographical locations and suitable for dealing with high-dimensional variables even for correlated variables.

In this study, we used two nonlinear regression models (RF, GW-RF) to identify the key risk factors to the COVID-19 death rate. The result showed that the nonlinear models effectively modelled the relationship between the risk factors and COVID-19 death rate both in global and local regressions. The adjusted R^2 of the GW-RF was 0.78, higher than that of the RF, indicating the GW-RF is more suitable to estimate the local risk factors of the COVID-19 death rate compared with the global model RF. The

282 average value of local R^2 of the GW-RF is 0.59. In GW-RF, the value of local R^2 higher than 0.4 in
283 89.4% of the counties and higher than 0.6 in 50.5% of the counties, indicating that the GW-RF performed
284 well in most of the study area. This shows that that the local nonlinear nonparametric model GW-RF can
285 accurately estimate the relationship between the risk factors and COVID-19 death rate at various
286 geographical locations.

287 Our result shows that several risk factors from environment, socioeconomic, demographic and
288 commuting to work are associated with the COVID-19 death rate. Finding of the global model RF
289 showed that householders with a mortgage had a highest correlation with the number of COVID-19 death
290 rate, followed by going to work by walking, land cover with forest, hospital beds, and overweight.
291 Findings of the geographical local model GW-RF is similar to that of the RF, but a little different. The the
292 GW-RF results show that going to work by walking, airborne benzene concentration, householder with a
293 mortgage, unemployment, airborne $PM_{2.5}$ concentration and percent of the black or African American
294 played an important role in the distribution of the COVID-19 death rate. Most of our findings are
295 consistent with previous research on COVID-19. Zheng et al. (2020) found that the frequency of public
296 transportation including flights, trains, and buses from the epicentre are important determinants of
297 transmission risks of COVID-19. They suggested preventive measures should be taken in public
298 transportation in order to contain the COVID-19 epidemic. Several studies found that the air pollution
299 have a significant correlation with the COVID-19 confirmed cases (Bashir et al. 2020; Xu et al. 2020).
300 Viruses are usually not spread as independent individuals in air, they are more likely to attach to other
301 suspended particles (Yang et al. 2011). Therefore, the concentration of air pollutants may affect the
302 aerosol transmission of SARS-CoV-2. These studies encouraged the formulation of environmental
303 policies to control pollution sources, which can reduce the harmful effects of air pollutants. Mollalo et al.
304 (2020) found that the proportion of black females and median household income had significant influence
305 on the spatial distribution of the COVID-19 incidence rate.

306 By exploring the spatial distribution of risk factors of the COVID-19 death rate, we found that
307 COVID-19 death rate in each region was affected by various factors and the association between each
308 risk factor and the COVID-19 death rate was not consistent in different spatial locations. About 35% of
309 the counties are most affected by Going to work by walking, so it is necessary to call on people to pay
310 attention to the social distancing and to wear medical masks. The western and central east region were
311 affected by the airborne benzene concentration; toxic particles in the air affect the spread of viruses.

Therefore, these regions should pay attention to the impact of air pollution on human health and take measures to protect the environment. The southern part of the continental United States was heavily affected by the proportion of the black or African American and householder with a mortgage, so there are some assistance probably can be taken in these regions to provide people with financial help such as food and medical supplies.

The current research, despite showing the spatial variability of the correlation between multiple risk factors and the COVID-19 death rate at a county level, has the following limitations. First, the current study only focused on the spatial dimension of the data based on a period, but the data about the COVID-19 death rate is constantly changing over time. Future study can research in the spatiotemporal direction. Secondly, we have not accounted for policy factors at local area. Policy factors would be an interesting research contribution to the transmission of COVID-19. Thirdly, the GW-RF model only assess the goodness-of-fit test of the regression, but does not assess the significance of the single variable. The test method of this model needs to be improved in the future study.

At present, few geographic local models study the nonlinear relationship between variables. The proposed GW-RF model could accurately estimate the spatial variability of nonlinear relationship between the risk factors and COVID-19 death rate, thus this method is applicable in many use instances where this is an issue about selecting significantly correlated variables at various geographical locations. Our results confirmed the findings of existing work on COVID-19, but extends it by using a nonlinear approach to quantify the impact of risk factors relevant in local areas. We expect this study could provide a reference for the geographical local nonlinear modelling in the future epidemiological studies.

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.

Credit authorship contribution statement

Yaowen Luo: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - Original Draft. Jianguo Yan: Conceptualization, Validation, Writing - Review & Editing, Supervision, Funding acquisition. Stephen McClure: Validation, Writing - Review & Editing.

Acknowledgements

This research is supported by a grant provided by National Scientific Foundation of China (Grant No.

U1831132 and 41374024), Innovation Group of Natural Fund of Hubei province (Grant No. 2018CFA087) and the fundamental research funds for the central universities (2042018KF0231).

Consent for publication

All the co-authors consent the publication of this work.

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.

Reference

Ahmadi M, Sharifi A, Dorosti S, Jafarzadeh Ghoushchi S, Ghanbari N (2020) Investigation of effective climatology parameters on COVID-19 outbreak in Iran. Science of The Total Environment 729:138705 doi:https://doi.org/10.1016/j.scitotenv.2020.138705

Bashir MF, Bilal BMA, Komal B (2020) Correlation between environmental pollution indicators and COVID-19 pandemic: A brief study in Californian context. Environmental Research:109652 doi:https://doi.org/10.1016/j.envres.2020.109652

Breiman L (2001) Random Forests vol 45. doi:10.1023/A:1010933404324

Brunsdon C, Fotheringham AS, Charlton ME (2010) Geographically Weighted Regression : A Method for Exploring Spatial Nonstationarity. geographical analysis 28:281-298 doi:10.1111/J.1538-4632.1996.TB00936.X

Cai H, Lam NSN, Qiang Y, Zou L, Correll RM, Mihunov V (2018) A synthesis of disaster resilience measurement methods and indices. international journal of disaster risk reduction 31:844-855 doi:10.1016/J.IJDRR.2018.07.015

Dariya B, Nagaraju GP (2020) Understanding novel COVID-19: Its impact on organ failure and risk assessment for diabetic and cancer patients. Cytokine & Growth Factor Reviews doi:https://doi.org/10.1016/j.cytogfr.2020.05.001

de León-Martínez LD, de la Sierra-de la Vega L, Palacios-Ramírez A, Rodriguez-Aguilar M, Flores-Ramírez R (2020) Critical review of social, environmental and health risk factors in the Mexican indigenous population and their capacity to respond to the COVID-19. Science of The Total Environment:139357 doi:https://doi.org/10.1016/j.scitotenv.2020.139357

Desjardins MR, Hohl A, Delmelle EM (2020) Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. Applied Geography 118:102202 doi:https://doi.org/10.1016/j.apgeog.2020.102202

Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.

Grömping U (2009) Variable Importance Assessment in Regression: Linear Regression versus Random Forest. the american statistician 63:308-319 doi:10.1198/TAST.2009.08199

Guliyev H (2020) Determining the spatial effects of COVID-19 using the spatial panel data model. Spatial Statistics 38:100443 doi:https://doi.org/10.1016/j.spasta.2020.100443

Huang C et al. (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet 395:497-506 doi:https://doi.org/10.1016/S0140-6736(20)30183-5

Imran M, Stein A, Zurita-Milla R (2015) Using geographically weighted regression kriging for crop yield mapping in West Africa. International Journal of Geographical Information Science 29:234-257 doi:10.1080/13658816.2014.959522

Ishwaran H (2007) Variable importance in binary regression trees and forests. electronic journal of statistics 1:519-537 doi:10.1214/07-EJS039

Landi F et al. (2020) THE GERIATRICIAN: THE FRONTLINE SPECIALIST IN THE TREATMENT OF COVID-19 PATIENTS. Journal of the American Medical Directors Association doi:https://doi.org/10.1016/j.jamda.2020.04.017

Lau H et al. (2020) The association between international and domestic air traffic and the coronavirus (COVID-19) outbreak. Journal of Microbiology, Immunology and Infection doi:https://doi.org/10.1016/j.jmii.2020.03.026

Liu F, Li X, Zhu G (2020) Using the contact network model and Metropolis-Hastings sampling to reconstruct the COVID-19 spread on the “Diamond Princess”. Science Bulletin doi:https://doi.org/10.1016/j.scib.2020.04.043

Lu B, Brunsdon C, Charlton M, Harris P (2017) Geographically weighted regression with parameter-specific distance metrics. International Journal of Geographical Information Science 31:982-998 doi:10.1080/13658816.2016.1263731

Malik VS, Ravindra K, Attri SV, Bhadada SK, Singh M (2020) Higher body mass index is an important risk factor in COVID-19 patients: a systematic review and meta-analysis. Environmental Science and Pollution Research doi:10.1007/s11356-020-10132-4

Marhl M, Grubelnik V, Magdič M, Markovič R (2020) Diabetes and metabolic syndrome as risk factors for COVID-19. Diabetes & Metabolic Syndrome: Clinical Research & Reviews doi:https://doi.org/10.1016/j.dsx.2020.05.013

Millett GA et al. (2020) Assessing Differential Impacts of COVID-19 on Black Communities. Annals of Epidemiology doi:https://doi.org/10.1016/j.annepidem.2020.05.003

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

O’Connor CM, Anoushiravani AA, DiCaprio MR, Healy WL, Iorio R (2020) Economic Recovery After the COVID-19 Pandemic: Resuming Elective Orthopedic Surgery and Total Joint Arthroplasty. The Journal of Arthroplasty doi:https://doi.org/10.1016/j.arth.2020.04.038

Rosenkrantz L, Schuurman N, Bell N, Amram O (2020) The need for GIScience in mapping COVID-19. Health & Place:102389 doi:https://doi.org/10.1016/j.healthplace.2020.102389

Ruthberg JS, Quereshy HA, Jella TK, Kocharyan A, D'Anza B, Maronian N, Otteson TD (2020) Geospatial analysis of COVID-19 and otolaryngologists above age

60. American Journal of Otolaryngology:102514 doi:https://doi.org/10.1016/j.amjoto.2020.102514
Serge R, Vandromme J, Charlotte M (2020) Are we equal in adversity? Does Covid-19 affect women and men differently? Maturitas doi:https://doi.org/10.1016/j.maturitas.2020.05.009
Shim E, Tariq A, Choi W, Lee Y, Chowell G (2020) Transmission potential and severity of COVID-19 in South Korea. International Journal of Infectious Diseases 93:339-344 doi:https://doi.org/10.1016/j.ijid.2020.03.031
Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) Conditional Variable Importance for Random Forests. bmc bioinformatics 9:307-307 doi:10.1186/1471-2105-9-307
Strobl C, Boulesteix A-L, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. bmc bioinformatics 8:25-25 doi:10.1186/1471-2105-8-25
Targher G et al. (2020) Patients with diabetes are at higher risk for severe illness from COVID-19. Diabetes & Metabolism doi:https://doi.org/10.1016/j.diabet.2020.05.001
Tosepu R, Gunawan J, Effendy DS, Ahmad LOAI, Lestari H, Bahar H, Asfian P (2020) Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. Science of The Total Environment 725:138436 doi:https://doi.org/10.1016/j.scitotenv.2020.138436
Wheeler D, Tiefelsdorf M (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. journal of geographical systems 7:161-187 doi:10.1007/S10109-005-0155-6
World Health Organization b (2020b) Coronavirus disease (COVID-19) outbreak situation. Accessed 21 March 2020
Xu H et al. (2020) Possible environmental effects on the spread of COVID-19 in China. Science of The Total Environment 731:139211 doi:https://doi.org/10.1016/j.scitotenv.2020.139211
Yang W, Elankumaran S, Marr LC (2011) Concentrations and size distributions of airborne influenza A viruses measured indoors at a health centre, a day-care centre and on aeroplanes. journal of the royal society interface 8:1176-1184 doi:10.1098/RSIF.2010.0686
Zheng R, Xu Y, Wang W, Ning G, Bi Y (2020) Spatial transmission of COVID-19 via public and private transportation in China. Travel Medicine and Infectious Disease 34:101626 doi:https://doi.org/10.1016/j.tmaid.2020.101626

Figures

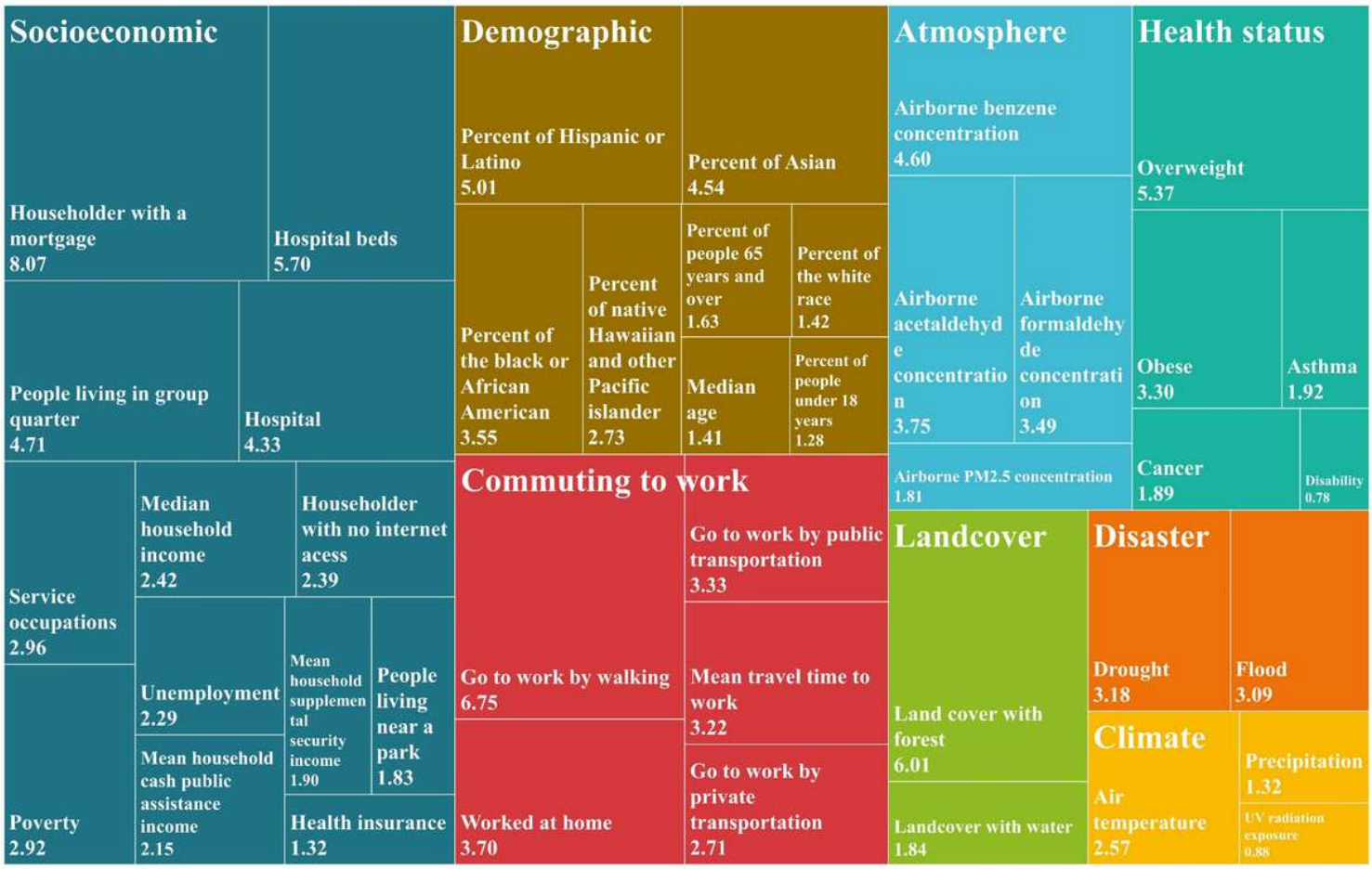

Figure 1

The variable importance of the independent variables of the RF model in modelling COVID-19 death rate

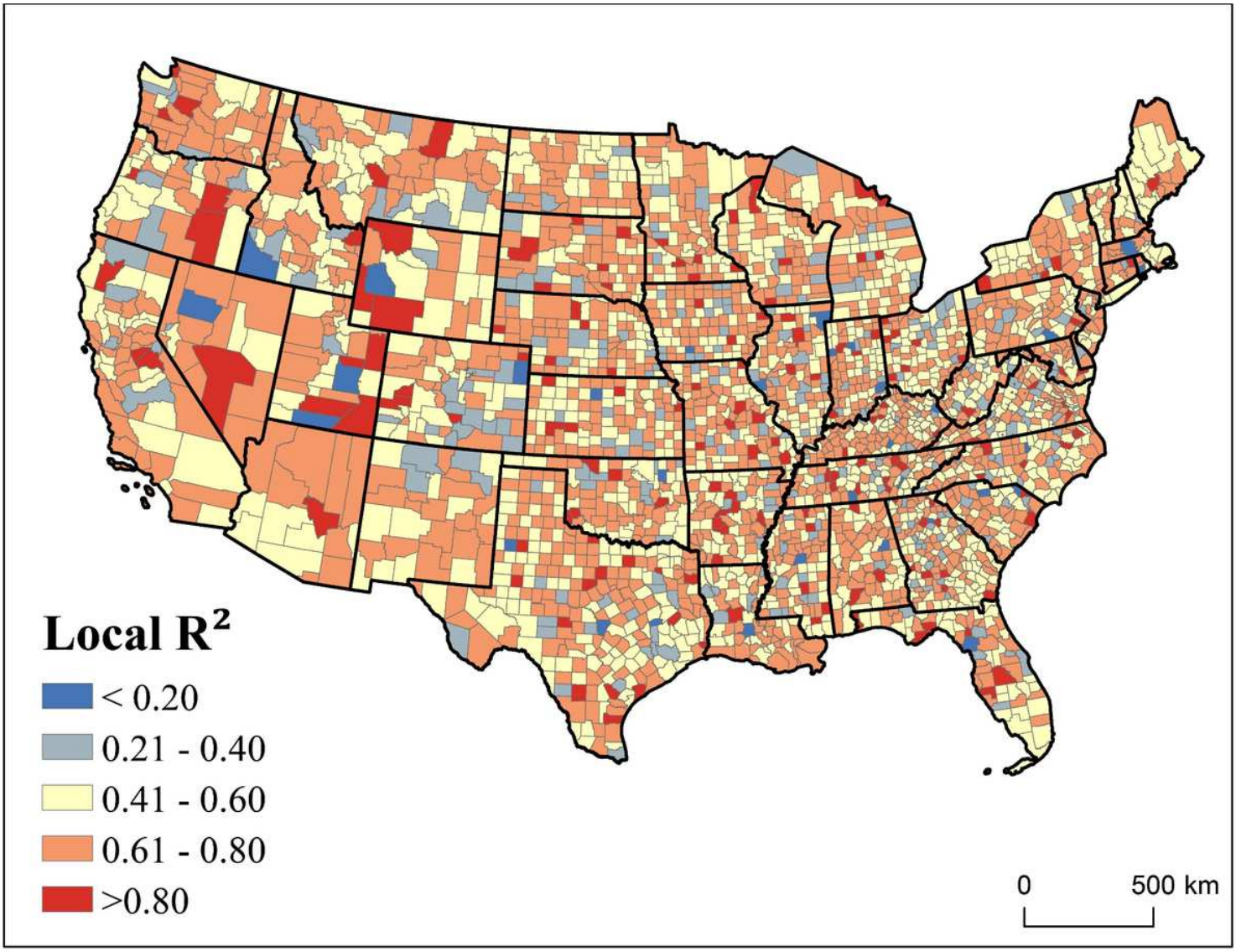

Figure 2

The distribution of local R^2 of the GW-RF

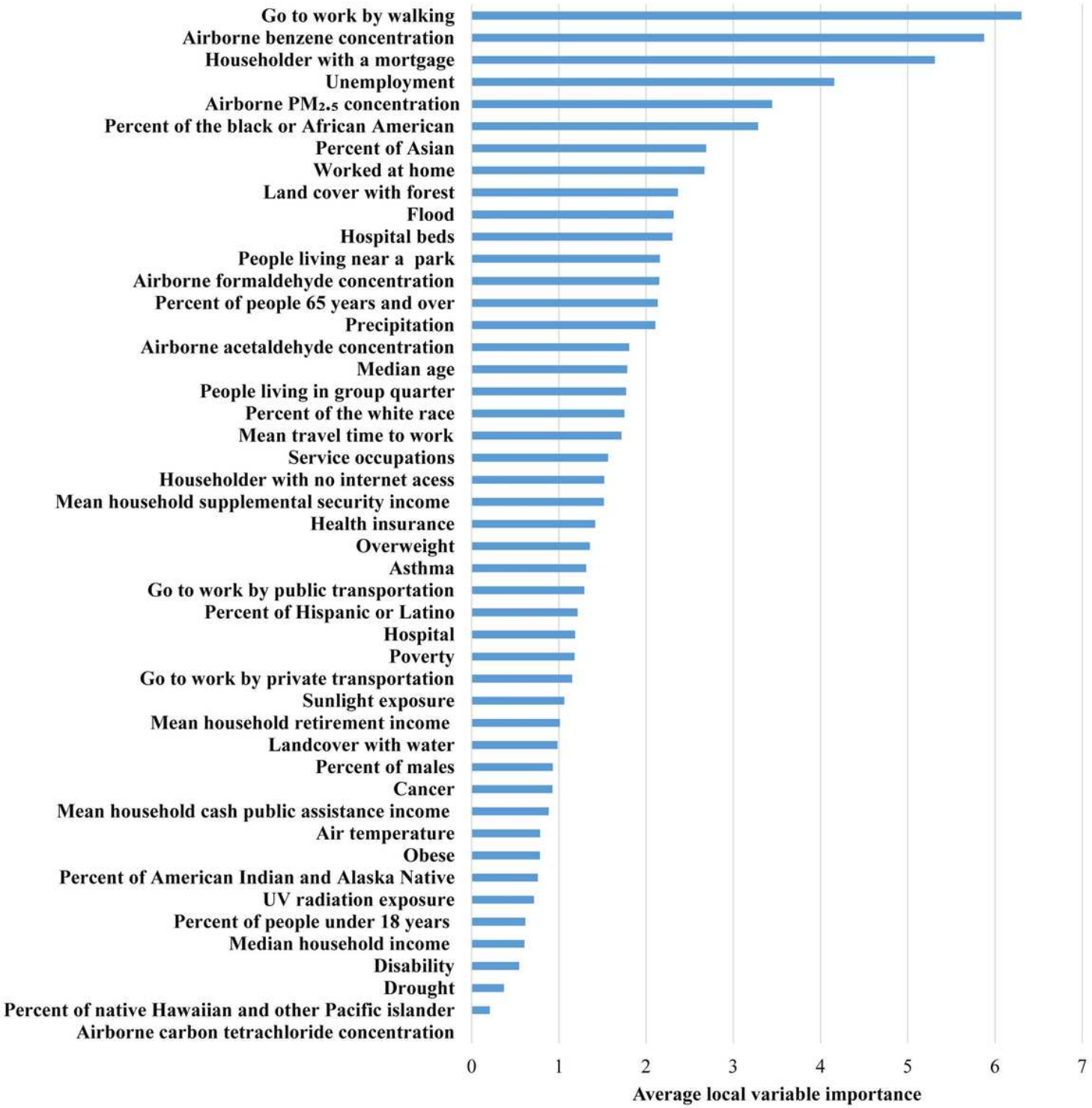

Figure 3

The average local variable importance of 47 potential risk factors on COVID-19 death rate in the GW-RF model

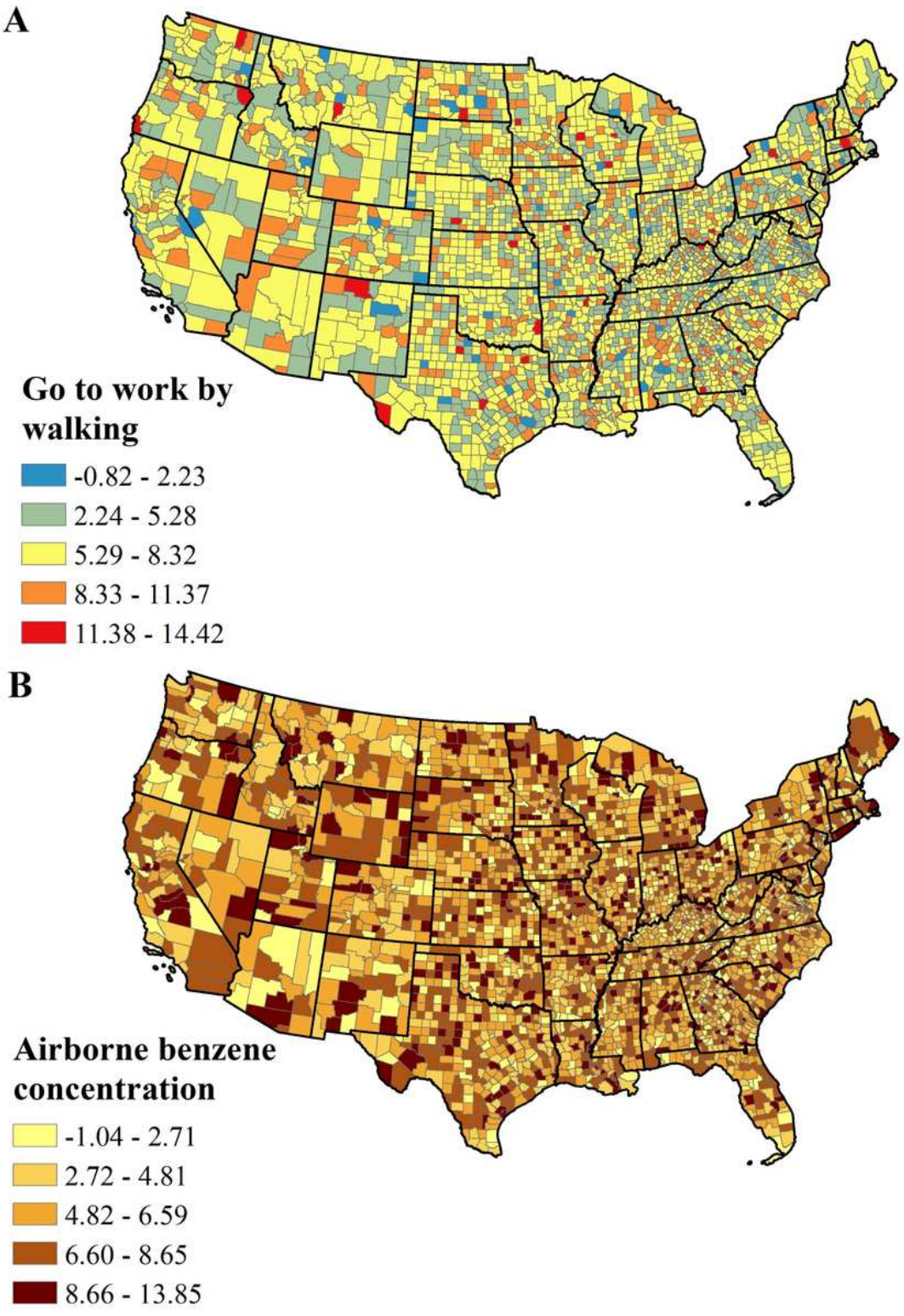

Figure 4

The spatial distribution of the local variable importance of (A) go to work by walking, (B) airborne benzene concentration on COVID-19 death rate in GW-RF mode

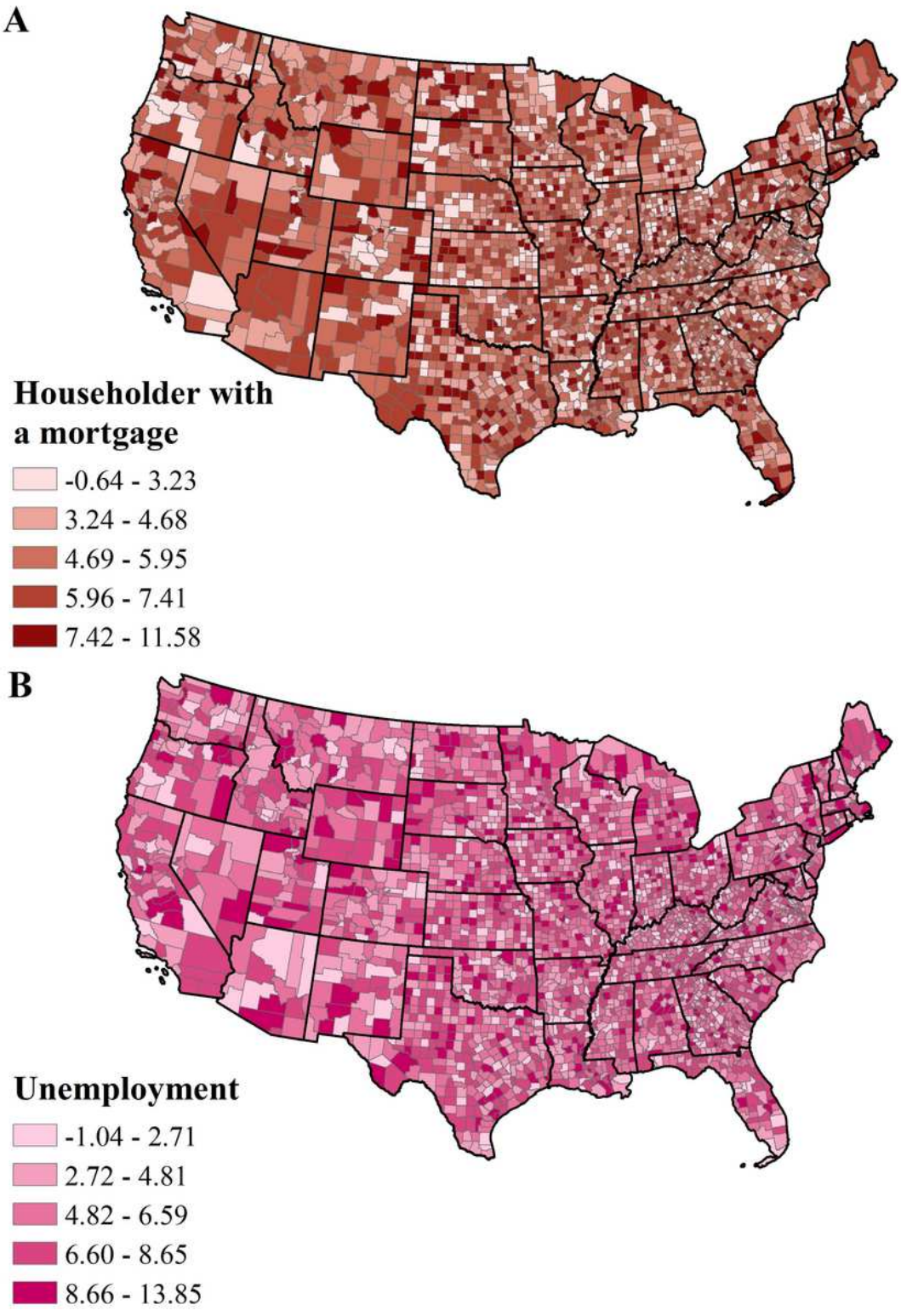

Figure 5

The spatial distribution of the local variable importance of (A) householder with a mortgage, (B) unemployment on COVID-19 death rate in GW-RF model

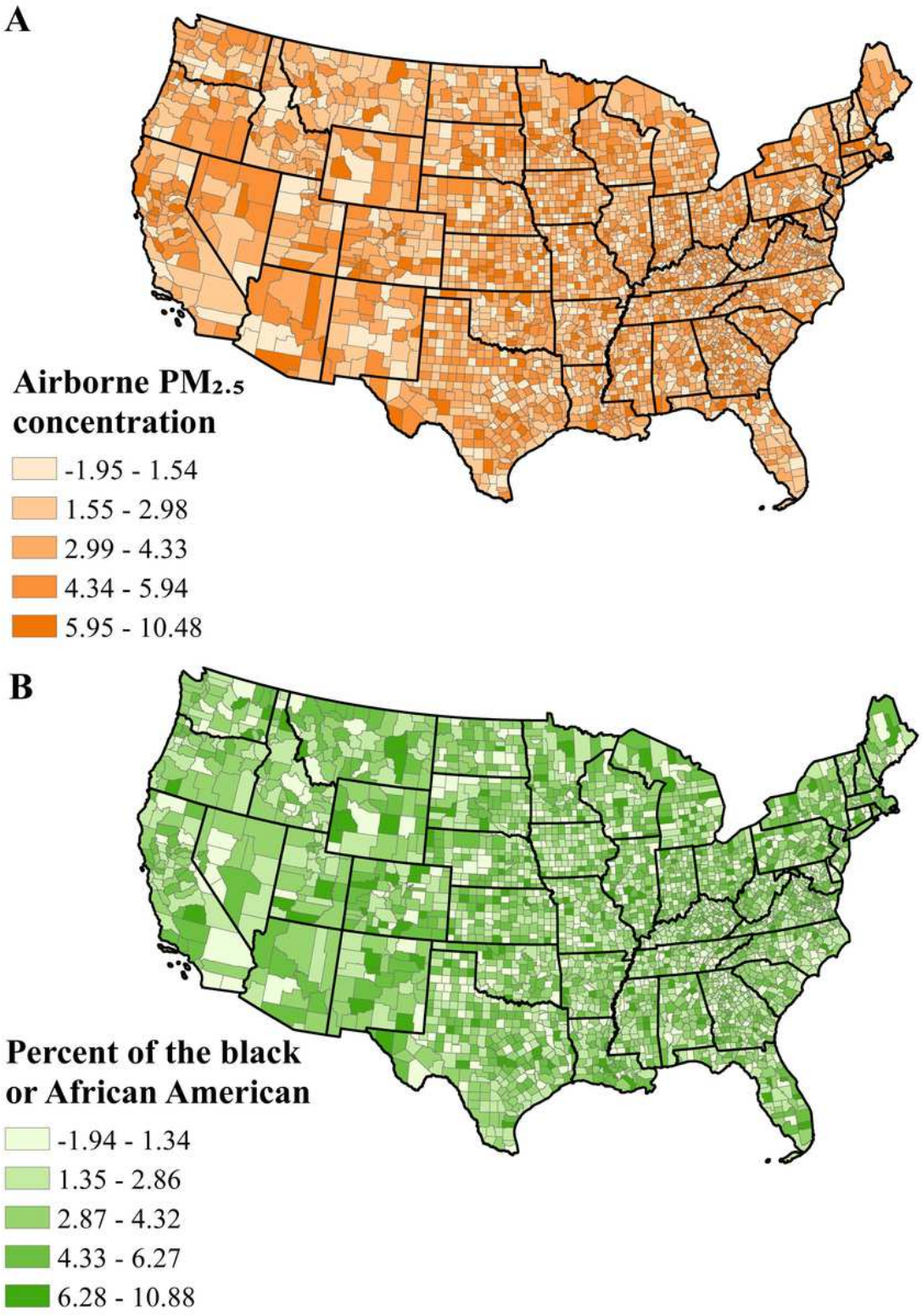

Figure 6

The spatial distribution of the local variable importance of (A) airborne PM 2.5 concentration, (B) percent of the black or African American on COVID-19 death rate in GW-RF model