Geographically varying relationships of COVID-19 mortality with different factors in India

Asif Iqbal Middya
Jadavpur University
Sarbani Roy (sarbani.roy@jadavpuruniversity.in)
Jadavpur University https://orcid.org/0000-0002-7598-8266

Research Article

Keywords: Covid-19, Geographically varying relationships, mortality, socioeconomic, environmental pollution, spatial relationships, correlation

Posted Date: October 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-93796/v1

Version of Record: A version of this preprint was published at Scientific Reports on April 12th, 2021. See the published version at https://doi.org/10.1038/s41598-021-86987-5.

Geographically varying relationships of COVID-19 mortality with different factors in India

Asif Iqbal Middya[1] **and Sarbani Roy**[1,*]

[1]Department of Computer Science and Engineering, Jadavpur University, Kolkata, 700032, India
[*]email: sarbani.roy@jadavpuruniversity.in

ABSTRACT

COVID-19 is a global crisis where India is going to be one of the most heavily affected countries. The variability in the distribution of COVID-19-related health outcomes might be related to many underlying variables, including demographic, socioeconomic, or environmental pollution related factors. The global and local models can be utilized to explore such relations. In this study, ordinary least square (global) and geographically weighted regression (local) methods are employed to explore the geographical relationships between COVID-19 deaths and different driving factors. It is also investigated whether geographical heterogeneity exists in the relationships. More specifically, in this paper, the geographical pattern of COVID-19 deaths and its relationships with different potential driving factors in India are investigated and analysed. Here, better knowledge and insights into geographical targeting of intervention against the COVID-19 pandemic can be generated by investigating the heterogeneity of spatial relationships. The results show that the local method (geographically weighted regression) generates better performance ($R^2 = 0.973$) with smaller Akaike Information Criterion (AICc $= -77.93$) as compared to the global method (ordinary least square). The GWR method also comes up with lower spatial autocorrelation (Moran's $I = -0.0436$ and $p < 0.01$) in the residuals. It is found that more than 87.5% of local R^2 values are larger than 0.60 and almost 60% of R^2 values are within the range $0.80 - 0.97$. Moreover, some interesting local variations in the relationships are also found.

Introduction

The novel coronavirus disease (COVID-19) has spread rapidly to all parts of the world, causing almost 1 million deaths as of mid-September 2020[1]. Because of its unpredictable nature and lack of appropriate medications, COVID-19 is now a global health concern. There is unprecedented urgency to investigate the major factors that are related to COVID-19 death. In this context, recent studies are focusing on exploring person-specific risk factors for COVID-19-related health outcomes[2–4]. Also, there are research works that examine the association of COVID-19-related health outcomes with different socio-economic, environmental, and region-specific factors[5–7]. These factors play a very important role in determining the patterns of COVID-19 mortality.

Both global and local models can be utilized to explore the above-mentioned associations. A global model comes up with a geographically constant relationship across the entire geographic space. On the other hand, a local model can capture the local relationships that can vary across the geographic space. Most of the studies that focus on exploring the relationship of COVID-19 cases with different possible risk factors are based on global models (e.g. Ordinary Least Square)[8,9]. But, the global models assume that the associations between the independent variables and the dependent variable are stationary (i.e. homogeneous) throughout the study area. Besides, these models also assume that there is no spatial autocorrelation in the dataset. Eventually, they yield estimates of the parameters that reflect average behaviour[10]. But, in reality, the relationships between the dependent and the independent variables may not be homogeneous and can be geographically varying[11]. Therefore, such models usually suffer from low accuracy especially in those locations where weak association exists between dependent and independent variables. Now, various local techniques can be utilized in order to overcome the above-mentioned shortcomings of the global models. Some widely encountered local spatial statistics include geographically weighted regression (GWR)[12,13], local Moran's I[14], spatial regressions, etc.

As of 17 September 2020, India is the world's second worst-affected country by COVID-19, with a total number of deaths exceeding 93000 thousand and a total number of confirmed cases exceeding 5.9 million[1]. However, in India, no comprehensive study is performed at the local level to investigate geographical relationships between COVID-19 deaths and associated potential factors. To bridge the gap, a local method (GWR) is employed to explore the geographical distribution and associated potential socio-economic, demographic, and environmental factors for COVID-19 deaths. Note that, the GWR model helps us to identify whether there is geographical heterogeneity present in the relationships. Moreover, a comparison between local (OLS) and global (GWR) models are also performed. This paper offers further knowledge and insight into geographical targeting of intervention and control strategies against the COVID-19 epidemic. In summary, the key objectives of this study are (i) to

explore the potential socio-economic, demographic, and environmental driving factors for COVID-19 deaths in India; (ii) to investigate geographically varying relationships of COVID-19 deaths with the driving factors by employing local (GWR) model. (iii) comparing the results of the local (GWR) model with the global (OLS) model to validate its suitability.

Materials and Methods

Data description

The geographical variabilities of COVID-19 deaths are modeled based on the district-level data across India. Note that, the COVID-19 mortality data are acquired for more than 400 districts in India. The geographical distributions of COVID-19 deaths are shown in Fig 1. The largest number of COVID-19 deaths are observed in the districts of the state Maharashtra. A total of 6 among 28 states contains at least one district that reports more than 1000 COVID-19 deaths. Table 1 summarizes all the raw datasets, their descriptions, the sources including the links from where these data can be found, and potential factors (independent variables) that are extracted from the raw datasets.

Datasets

Three raw datasets are mainly utilized to investigate geographically varying relationships of COVID-19 deaths with different environmental, demographic, and socio-economic factors. The first dataset includes district wise COVID-19 death counts in India. The cumulative number of COVID-19-related deaths for each district is collected up to September 8, 2020, from the COVID19INDIA website (https://www.covid19india.org/). COVID19INDIA is a crowdsourced initiative to document the COVID-19 data from the states and union territories of India. The second dataset pertaining to environmental pollution includes the daily concentration of different air pollutants (e.g. $PM_{2.5}$, SO_2, NO_2, etc.). The concentration of air pollutants (from January 2016 to January 2020) for a total of 130 monitoring stations are obtained from the Central Pollution Control Board (CPCB[15]), INDIA. The third dataset contains socio-economic and demographic data that may have an association with COVID-19 mortality. The district-level socio-economic and demographic data are obtained from the last census in India that was conducted in 2011.

Additionally, the district-level data of each district needs to be linked with the GPS coordinate of the centroid of that district. The dataset containing GPS coordinates of the districts of India are collected from Kaggle (https://www.kaggle.com/sirpunch/indian-census-data-with-geospatial-indexing).

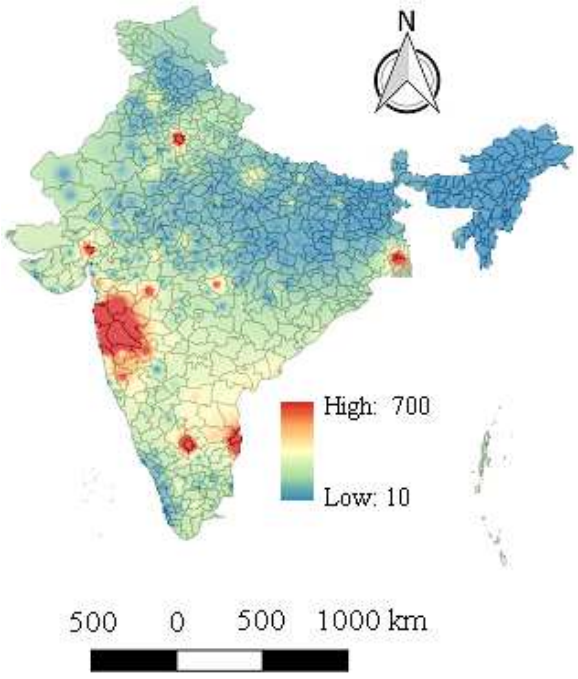

Figure 1. Geographical distribution of COVID-19 deaths across India. The spatially continuous distribution map is generated in QGIS (https://qgis.org/en/site/) by using Inverse Distance Weighting (IDW) interpolation.

Table 1. A summary of datasets

Dataset	Dataset description	Source	Variable name	Variable Explanation
COVID-19 data	COVID-19 data from the states and union territories of India up to September 8, 2020.	(i) COVID19INDIA website (https://www.covid19india.org/) (ii) Ministry of Health and Family Welfare, Government of India (https://www.mohfw.gov.in/)	*COVID19_Death*	District-level COVID-19 death count up to September 8, 2020
Census data	Contains district wise socioeconomic and demographic data of India	India census, 2011 (https://censusindia.gov.in/)	*Tot_Population*	District-level total population
			HH_Above_8_P	District-level count of total number of households with at least 9 persons
			Growth_Rate	District-level rate at which population increases.
			Sex_Ratio	District-level count of the number of females per 1000 males
			Age_Abv_50	District-level count of total number of persons with age 50 years or more
			HH_With_TCMC	District-level count of total number of households having TV, Computer (or laptop), Mobile phone, and Car.
			Higher_Edu	District-level count of total number of persons having higher education
			P_Urb_Pop	District level percentage of urban population
Environmental air pollution data	Concentration of air pollutants (from January, 2016 to January, 2020) for a total of 202 monitoring stations	(i) Central Pollution Control Board (CPCB), Government of India (https://cpcb.nic.in/)	$PM_{2.5}$	District-level exposure to $PM2.5$, averaged across the period 2016 – 2020
			NO_2	District-level exposure to NO_2, averaged across the period 2016 – 2020
			SO_2	District-level exposure to SO_2, averaged across the period 2016 – 2020

Table 2

Variable	VIF
Tot_Population	7.92
Growth_Rate	1.24
Sex_Ratio	1.30
HH_With_TCMC	3.91
HH_Abv_8_P	12.4
Higher_Edu	3.12
$PM_{2.5}$	2.65
Age_Abv_50	3.47
P_Urb_Pop	1.93
SO_2	1.20
NO_2	1.52

Table 3

Variable	VIF
Tot_Population	3.93
Growth_Rate	1.21
Sex_Ratio	1.32
HH_With_TCMC	2.96
Higher_Edu	2.94
$PM_{2.5}$	2.35
Age_Abv_50	2.87
P_Urb_Pop	1.80
SO_2	1.19
NO_2	1.54

Data preparation

From the raw datasets, a total of eleven potential demographic, socioeconomic, and environmental pollution related factors (see Table 1) are selected to explain the district-level geographical variation of COVID-19 mortality. The district-level demographic and socioeconomic factors that are selected in this study are: population; households with at least 8 persons; growth rate; sex ratio; persons with age 50 years or more; households having TV, computer (or laptop), mobile phones and car; number of persons having higher education; the percentage of the urban population. On the other hand, the environmental pollution related variables that are selected are as follows: $PM_{2.5}$ exposure; NO_2 exposure; SO_2 exposure.

The district-level long-term exposure to three air pollutants namely $PM2.5$, NO_2, and SO_2 are calculated from the raw data of 202 pollution monitoring stations. The mean concentration of each of the above-mentioned air pollutants of all the 202 monitoring stations is computed for the period 2016-2020. For each pollutant, the computed values are spatially aggregated by averaging the values of all monitoring stations of a district. If a district doesn't contain any monitoring stations, then it's exposure to that pollutant is computed using Nearest Neighbour interpolation (NNI).

A multicollinearity verification is performed via the Variance Inflation Factor (VIF) to remove unnecessary redundancy

among the 10 explanatory variables. VIF can be expressed as follows:

$$VIF^k = \frac{1}{1-R_k^2} \tag{1}$$

where, R_k^2 denotes the coefficient of determination that is computed by regressing the k^{th} variable on remaining explanatory variables. Firstly, regression analysis is conducted among all the 11 explanatory variables to compute the VIFs that are shown in Table 2. It is observed that the variable *HH_Abv_8_P* has high Variance Inflation Factor (VIF=12.4). Now, if VIFs are larger than 10, it indicates that there is multicollinearity[16]. Eventually, the variable *HH_Abv_8_P* is removed from the set of explanatory variables. After that, the regression is again performed on the remaining 10 variables, with the VIFs given in Table 3. Now, it is observed that no VIF exceeds 10 eventually this set of 10 variables can be used for model building.

Modeling spatial relationship

In this paper, the OLS (Ordinary Least Square) and GWR (Geographically Weighted Regression) models are utilized to determine the geographical relationship of COVID-19 mortality with potential risk factors.

The OLS method generally attempts to understand the global relationships between the dependent and independent variables. In this case, the regression and its parameters are unchanged over the geographic space. Mathematically, Eq. 2 represents a global regression model as follows:

$$\mathcal{Y}_i = \eta_0 + \sum_{k=1}^{n} \eta_k \mathcal{X}_{ik} + \delta_i \tag{2}$$

where, $\mathcal{Y}_i$ denotes the dependent or response variable; $\mathcal{X}_{ik}$ is the i^{th} observation of k^{th} independent variable; η_k the global regression coefficient for k^{th} independent variable; η_0 represent the intercept parameter; and δ_0 denotes the error term.

GWR technique extends the global regression (Eq. 2) by enabling local parameter estimation[13]. It allows regression coefficients to be a function of geographical location. In other words, the regression coefficients are quantified independently in different geographical locations. A GWR model (Eq. 3) can be represented as follows:

$$\mathcal{Y}_i = \xi_{i0} + \sum_{k=1}^{n} \xi_k(\mu_i, \nu_i)\mathcal{X}_{ki} + \delta_i \tag{3}$$

where, $\mathcal{Y}_i$, $\mathcal{X}_{ki}$, and δ_i denote the dependent (or response) variable, k^{th} independent (or predictor) variable, and error at location i respectively; (μ_i, ν_i) denotes coordinates of location i; $\xi_k(\mu_i, \nu_i)$ represent local coefficient for k^{th} predictor at location i. Note that, GWR model allows regression parameters to vary continuously across the geographic space. For each location i, a set of regression parameters is estimated. The estimation of parameters can be performed as follows:

$$\widehat{\xi}(\mu,\nu) = (\mathscr{X}^T \mathscr{W}(\mu,\nu)\mathscr{X})^{-1}\mathscr{X}^T \mathscr{W}(\mu,\nu)\mathcal{Y} \tag{4}$$

where, $\mathscr{X}$ denotes a matrix containing the values of independent variables and a column of all 1s; $\mathcal{Y}$ represents a vector of values of the dependent variable; $\widehat{\xi}(\mu,\nu)$ is a vector of local regression parameters; $\mathscr{W}(\mu,\nu)$ is a diagonal matrix whose diagonal elements represent the geographical weighting of the observations for regression location. The weights in $\mathscr{W}(\mu,\nu)$ assigns greater weights to the observations that are closer to the regression point than the observations that are farther away. In this work, the weights are computed using a Gaussian kernel function which is defined as follows:

$$\begin{cases} w_{ij} = exp\Big[-\frac{1}{2}\Big(\frac{D_i^j}{B}\Big)^2\Big], & \text{if } D_i^j \le B \\ w_{ij} = 0, & \text{otherwise} \end{cases} \tag{5}$$

where, B represents the bandwidth and D_i^j denotes the distance between the regression point i and the location of observation j. Note that, the bandwidth can be defined either by a fixed number of closest neighbors (known as adaptive bandwidth) or by a fixed distance (known as fixed bandwidth). Golden Section search[17] is utilized to find the optimum size of the bandwidth for GWR.

Performance metrics

The performance of the models are assessed by three metrics namely R^2, adjusted R^2, and AICc. Here, AICc is a corrected version of the Akaike Information Criterion (AIC). AICc can be defined as follows[13]:

$$AICc = N\ ln(2\pi) + 2N\ ln(\hat{\sigma}) + N \times \left(\frac{N + tr(S)}{N - 2 - tr(S)}\right) \tag{6}$$

where, N denotes the sample size, S is the hat matrix, $tr(S)$ denotes the trace of S, and $\hat{\sigma}$ represents the estimated standard deviation of the error term. AICc denotes model's accuracy and lower AICc indicates better model quality. It is usually used to find the best-fit model. The value of R^2 represents the ability of a model to explain the variance in the dependent variable and therefore a larger R^2 signifies the better performance of the model. It is computed from the estimated and the actual values of the dependent variable. Moreover, Moran's I index is computed to investigate the spatial autocorrelation of the model residuals. Mathematically, it is defined as follows:

$$I = \frac{N\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\right)\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{7}$$

where, N denotes total number of observations, y_i and y_j are variable values at location i and j respectively,$\bar{y}$ represents the mean value, and w_{ij} denotes a weight between location i and j. The value of Moran's I index can vary between -1 (perfect dispersion) to +1 (a perfect positive autocorrelation)). Note that, a zero value indicates perfect spatial randomness.

Model Building

Here, a step-wise GWR model selection using AICc is presented that can be utilized to investigate geographically varying relationships of COVID-19 mortality with different driving factors. The following are the steps to build an appropriate GWR model[18].

- *Step 1:* Suppose there are n explanatory variables (in our case $n = 10$). For each of the explanatory variables, fit a separate GWR model by regressing that variable against the $COVID19\_Death$ variable. Compute AICc for each of the $n = 10$ models. Find the model that generates the lowest AICc and permanently include the corresponding explanatory variable in subsequent model building.

- *Step 2:* Subsequently select a variable from the remaining $(n-1)$ variables, build a model with the permanently included variables along with the newly selected variable. Find the explanatory variable that produces the lowest AICc and permanently include it in subsequent model building. Set $n = n-1$.

- *Step 3:* Repeat Step 2 until it is observed that there is no reduction in AICc.

The above-mentioned steps are carried out using MGWR 2.2 software[19]. When calibrating the GWR, an adaptive bisquare spatial kernel is applied. Moreover, in order to select an optimal bandwidth, the Golden Section search[17] is employed. Fig. 2 shows the changes in AICc during the step-by-step selection of explanatory variables for model building. It is observed that after the inclusion of a total of five variables, the AICc values start increasing when further new variables are included. Note that, both a global (OLS) model and a local (GWR) model are calibrated with these five explanatory variables.

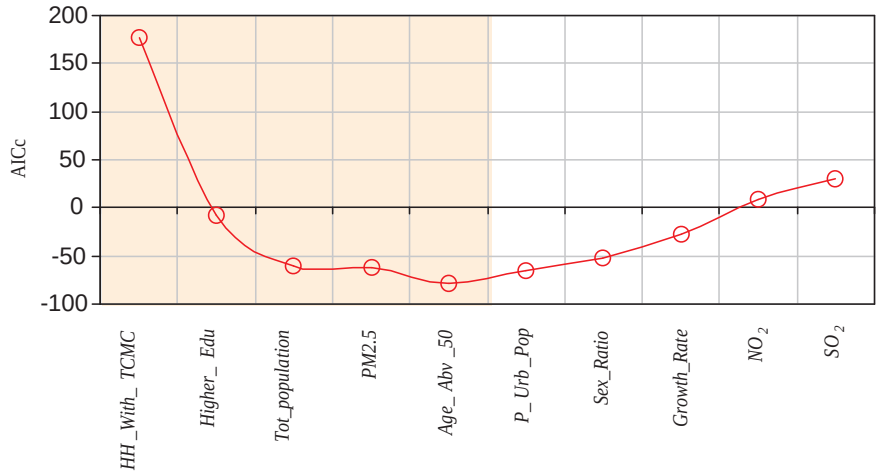

Figure 2. Stepwise variable selection for geographically weighted regression (GWR).

Results

In this section, firstly the performance of the global model (OLS) and local model (GWR) are discussed. Next, the geographically varying relationships of COVID-19 mortality with different factors are presented.

Performance of OLS and GWR Model

A detailed summary of the OLS model is presented in Table 4. The variables *Tot_population*, *HH_With_TCMC*, *Age_ABV_50*, and $PM_{2.5}$ returns significant t values of 2.758, 11.637, -1.868 and -2.509 respectively. Moreover, the Moran's I of the residuals of the global OLS model are also analysed. It is found that there is significant spatial autocorrelation (Moran's I = 0.35 and p ¡ 0.01). The assumptions of OLS estimation are violated as there exist dependent residuals. Eventually, the GWR model is utilized to show the geographical variations of the relationships with different factors. A detailed summary of the GWR model for the local parameter estimates is presented in Table 5.

Table 4. Summary of the global model (OLS) for various socioeconomic, demographic, and environmental pollution related factors.

Variable	Coef. Est	Est Err	t statistic	p-value
Intercept	0.000	0.032	0.000	1.000
Tot_population	0.301	0.109	2.758	0.006*
HH_With_TCMC	0.673	0.058	11.637	0.000*
Age_Abv_50	-0.237	0.127	-1.868	0.050*
Higher_Edu	0.062	0.095	1.024	0.306
$PM_{2.5}$	-0.097	0.039	-2.509	0.012*

* Significant at 0.05

Table 5. Summary of the local model (GWR) for various socioeconomic, demographic, and environmental pollution related factors.

Variable	Mean	STD	Min	Median	Max
Intercept	0.066	0.264	-0.210	-0.059	0.879
Tot_population	0.111	0.428	-1.371	0.073	0.773
HH_With_TCMC	0.384	0.356	-0.332	0.290	2.194
Age_Abv_50	0.259	0.450	-0.612	0.143	2.305
Higher_Edu	-0.019	0.574	-1.501	-0.063	2.026
$PM_{2.5}$	0.089	0.241	-0.155	0.010	1.201

The performance of OLS and GWR model in terms of R^2, $Adj\ R^2$, and AICc are also provided in Table 6. It is observed that the GWR model resulted in a better fit as compared to the global OLS model. The global model explains only 65.8% of the variance of district-level COVID-19 deaths which is increased to 97.3% if the model is calibrated as GWR by taking into account the local impact of the explanatory variables. Comparing the models in terms of AICc, show that the model fit is greatly enhanced by reducing the value of AICc from 605.494 (OLS model) to -77.936 (GWR model). Moreover, the verification of Moran's I of the residuals of the GWR model indicates that the residuals are randomly distributed (Moran's I=-0.0436 and p ¡ 0.01). In other words, the residuals don't have any significant spatial autocorrelation and eventually, it shows the suitability of GWR over the global model (OLS).

Geographically varying relationships between COVID-19 deaths and the driving factors

The geographical distribution of R^2 is presented in Fig. 3 that shows it varies within a range $0.38 - 0.97$. It is found that more than 87.5% of local R^2 values are larger than 0.60 and almost 60% of R^2 values are within the range $0.80 - 0.97$. Note that, very high R^2 values are mainly observed in the western and the eastern regions of India. Moreover, low and moderate R^2 values are mainly distributed over the northern and the southern part of India.

Table 6. Performance comparison OLS and GWR models in terms of three performance metrics: AICc, R^2, and Adjusted R^2.

Performance metrics	OLS	GWR
AICc	605.494	-77.936
R^2	0.658	0.973
$Adj\ R^2$	0.653	0.966

Now, the geographical distribution of local coefficient estimates of the GWR model is provided in Fig. 4 to further reveal the relationship of the explanatory variables with the COVID-19 deaths. It mainly facilitates understanding of the complex relationship that varies over the geographic space. The results of GWR in Fig. 4 not only present positive or negative relationships but also show whether the relationship is strong or weak. A positive relationship indicates that the COVID-19 deaths tend to increase as the value of specific explanatory variable increases. A negative relationship indicates that the COVID-19 deaths tend to decrease as the value of specific explanatory variable increases. Moreover, larger values of a coefficient denote a stronger relationship. In the maps of Fig. 4, the regions having deep red shade denote regions in which the specific variable has a strong positive influence (i.e. strong positive relationship) on COVID-19 deaths.

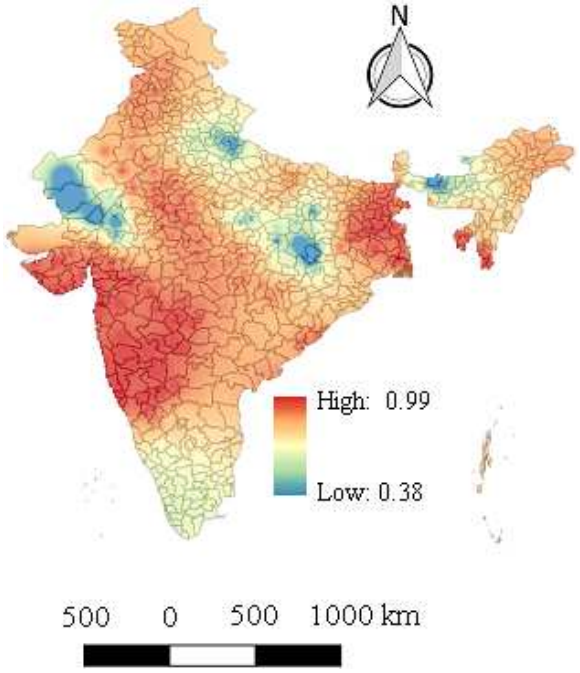

Figure 3. Geographical distribution of R^2 values for geographically weighted regression (GWR) model.

As shown in Fig. 4 (a), the GWR model produces local intercept that can vary within the range -0.21 to 0.87 with a mean of 0.06. In Fig. 4 (b), the regions with deep red color (mainly the state of West Bengal) denote those areas where the variable *HH_With_TCMC* has a strong positive relationship with COVID-19 death. The variable *Higher_Edu* is a strong predictor (See Fig. 4 (c)) for COVID-19 death in some parts of western India (mainly the state of Gujarat), southern India (mainly the states of Tamil Nadu and Kerala), and Eastern India (mainly the state of West Bengal). On the other hand, in the southern and the south-western part of India, a positive relationship between population and COVID-19 death is found (see Fig. 4 (d)). However, in some regions of central and western India (the states of Madhya Pradesh and Gujarat), a strong negative relationship between population and COVID-19 death is also observed. Fig. 4 (e) shows that mainly in the western part of India there is a strong positive relationship between $PM_{2.5}$ and COVID-19 death, whereas in the other parts of India there is no such strong relationship. The explanatory variable *Age_Abv_50* shows a positive relationship in central, eastern, and northern parts of India (see Fig. 4 (f)).

Moreover, Table 7 represents the district-level results of the local model (GWR) for some of the districts that are severely affected by COVID-19 disease. It is observed that the GWR model yields high local R^2 value for most of the heavily affected districts. For instance, very high local R^2 values are found for the following districts: Pune ($R^2 = 0.987$), Thane

($R^2 = 0.988$), Nasik ($R^2 = 0.987$), Solapur ($R^2 = 0.989$), Kolhapur ($R^2 = 0.990$), Sangli ($R^2 = 0.991$), Satara ($R^2 = 0.990$), Dharwad ($R^2 = 0.980$), Latur ($R^2 = 0.984$), Mumbai ($R^2 = 0.972$), Kolkata ($R^2 = 0.978$), Chennai ($R^2 = 0.969$), Jalgaon ($R^2 = 0.968$), Nanded ($R^2 = 0.966$). On the other hand, moderate local R^2 values are found for Lucknow ($R^2 = 0.942$), Jaipur ($R^2 = 0.916$), Srikakulam ($R^2 = 0.926$), Ludhiana ($R^2 = 0.860$), Guntur ($R^2 = 0.828$), Kurnool ($R^2 = 0.837$), West Godavari ($R^2 = 0.862$), Bhopal ($R^2 = 0.89$), and Krishna ($R^2 = 0.844$). The lowest R^2 values are observed for the following districts: Chittor ($R^2 = 0.736$), Anantapur ($R^2 = 0.755$), Coimbatore ($R^2 = 0.608$), Hassan ($R^2 = 0.664$). Note that, for most of the highly COVID-19-affected districts, the variables $PM_{2.5}$ and HH_With_TCMC are usually exhibited positive relationships in regression modeling. On the other hand, the variable $Higher\_Edu$ usually exhibits negative relationships for most of the highly affected districts.

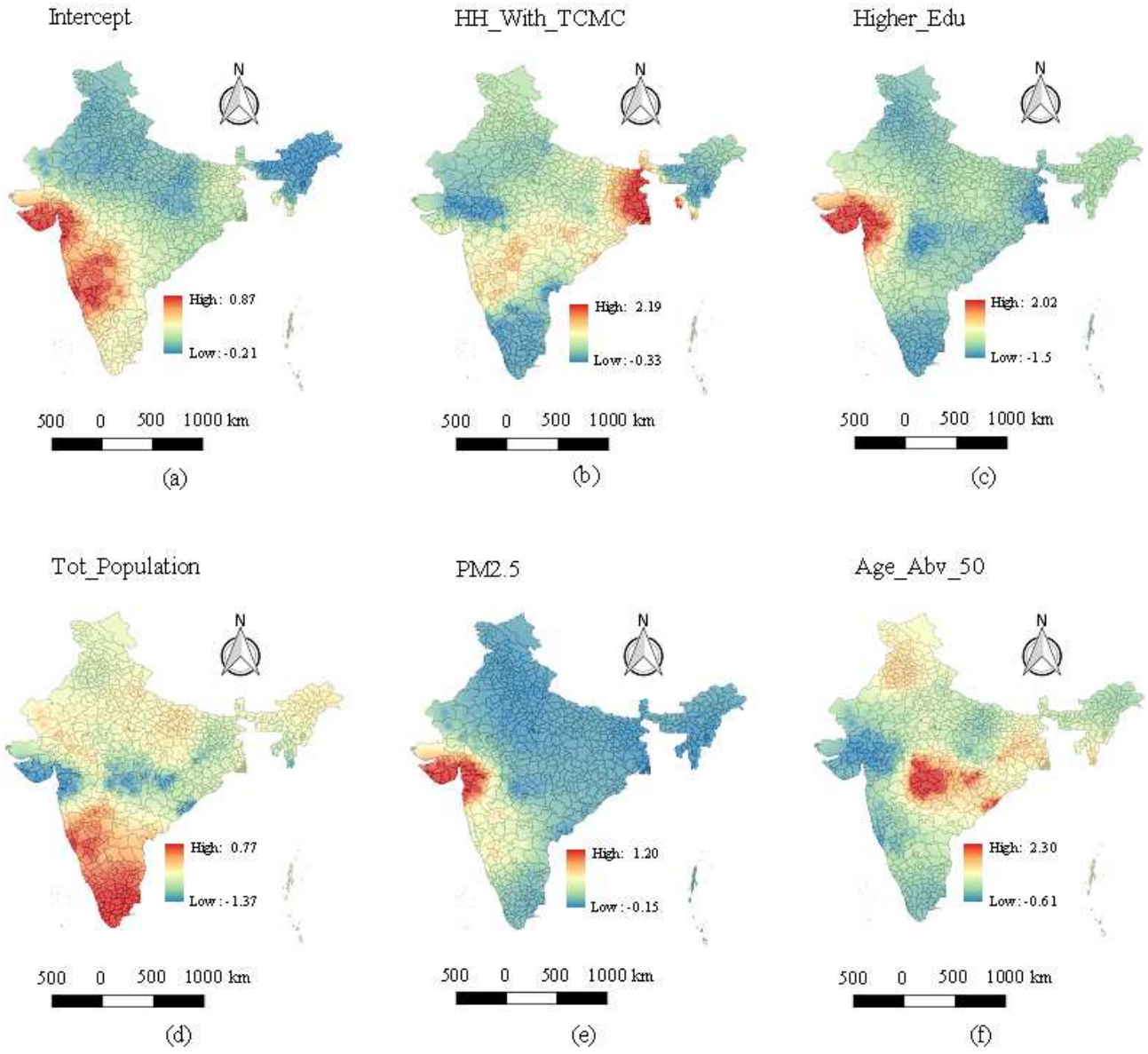

Figure 4. Local parameter estimates of geographically weighted regression (a) Intercept (b) HH_With_TCMC (c) $Higher\_Edu$ (d) $Tot\_population$ (e) $PM2.5$ and (f) Age_Abv_50.

Discussion

In order to better understand how different driving factors influence the overall fatalities caused by COVID-19, the geographical distribution of COVID-19-related deaths are investigated. The highest number of COVID-19-related deaths are found primarily

Table 7. Local R^2 and district-level parameter estimates by geographically weighted regression for some of the districts of India that are severely affected by COVID-19 disease.

District	Local R^2	Parameter estimates					
		Intercept	*PM*2.5	*Tot_population*	*HH_With_TCMC*	*Age_Abv_*50	*Higher_Edu*
Pune	0.987	-0.172	-0.390	-0.210	0.768	-0.146	0.878
Mumbai	0.972	0.791	0.665	-0.531	0.954	-0.093	0.765
Thane	0.988	0.648	0.911	-1.141	0.692	-0.075	1.709
Chennai	0.969	0.868	0.341	0.650	1.162	-0.56	-0.690
Kolkata	0.978	0.044	-0.048	-0.153	2.193	0.739	-1.500
Nashik	0.986	0.608	0.914	-1.280	0.647	0.089	1.748
Jalgaon	0.968	0.260	0.321	-0.465	0.616	0.302	0.742
Nagpur	0.881	0.123	-0.028	-1.104	0.741	1.867	-0.636
Solapur	0.989	0.650	0.436	0.496	0.751	-0.043	0.188
Kolhapur	0.990	0.551	0.279	0.539	0.719	-0.151	0.294
Surat	0.978	0.794	1.130	-1.269	0.520	-0.084	1.961
Sangli	0.991	0.591	0.349	0.487	0.727	-0.107	0.290
Ludhiana	0.860	-0.105	-0.013	-0.252	0.363	0.621	-0.363
Chittoor	0.736	0.127	0.063	0.474	0.044	0.102	-0.383
East Godavari	0.974	0.070	-0.250	-0.310	1.179	0.619	-0.731
Indore	0.807	0.078	0.074	0.188	-0.056	-0.134	0.633
Guntur	0.828	-0.002	0.031	0.266	-0.225	0.224	-0.070
Lucknow	0.942	-0.076	-0.014	0.087	0.258	-0.168	0.227
Satara	0.990	0.503	0.330	0.251	0.721	-0.127	0.517
Kurnool	0.837	0.551	0.354	0.271	0.327	0.054	0.082
Madurai	0.613	0.190	0.035	0.655	0.126	-0.076	-0.466
Anantapur	0.755	0.396	0.297	0.185	0.011	0.280	-0.192
Dharwad	0.980	0.685	0.394	0.612	0.780	-0.126	0.047
West Godavari	0.862	0.044	-0.046	0.043	0.296	0.349	-0.230
Coimbatore	0.608	0.203	0.055	0.615	0.115	-0.043	-0.462
Prakasam	0.818	0.055	0.088	0.263	-0.304	0.232	-0.059
Bhopal	0.894	-0.026	0.020	-0.259	0.402	0.300	0.045
Krishna	0.844	0.008	-0.010	0.188	0.031	0.260	-0.141
Latur	0.984	0.653	0.402	0.428	0.858	0.440	-0.380
Jaipur	0.916	-0.166	-0.027	0.001	0.214	0.112	-0.145
Srikakulam	0.926	-0.018	0.042	-1.160	0.908	1.349	-0.378
Nanded	0.966	0.414	0.073	-0.023	0.905	1.182	-0.806
Dhule	0.971	0.469	0.761	-0.894	0.459	-0.119	1.679
Hassan	0.664	0.319	0.198	0.367	0.020	0.157	-0.365

in the western part of India (Pune, Thane, Mumbai, Nagpur, Nashik, Raigad, Jalgaon, Kolhapur, Sangli, Satara, Solapur, Ahmedabad, Surat). On the other hand, the number of COVID-19-related deaths is relatively low in the northern and eastern parts of India. This study identified considerable geographical variability of COVID-19 deaths and their heterogeneous relationship at the local level with the driving factors in India. More specifically, the utilization of the GWR method successfully found the geographically varying relationship of COVID-19 mortality with various potential socio-economic, demographic, and environmental pollution related factors. This study reveals five important local factors are significantly related with district-level COVID-19 deaths as follows: (i) population (ii) $PM_{2.5}$ level (iii) households having TV, computer (or laptop), mobile phones and car (iv) persons with age 50 years or more (v) number of persons having higher education. Furthermore, this study also validates the effectiveness of local parameter estimation by comparing the global OLS method with the local GDR method. To the best of our knowledge, this is the first study that explores geographically varying relationships of COVID-19 deaths with various potential driving factors in India.

Rigorous analyses are performed to demonstrate the shortcomings of global technique (OLS) as compared to the local

technique (GWR) in terms of several performance metrics. The OLS model only explains 65.8% of the variance of district-level COVID-19 deaths. It is found that the predictive efficiency and model accuracy are further enhanced by implementing the GWR method. The GWR model explains 97.3% of the variance of district-level COVID-19 deaths. Moreover, Moran's I index verifies that no significant spatial autocorrelation is present in the residuals of the GWR model. Note that, a key advantage of such a local method is its capability to visualize the geographically varying heterogeneous relationships between the dependent and the independent variables. In other words, it enables us for a better understanding of relationships based on geographical contexts and study area's known features.

The findings of this study reveal that there are strong positive relationships of COVID-19 deaths with the explanatory variables $PM2.5$ and $Tot\_population$ across the regions of the COVID-19 death hotspots in the western part of India. The positive association of COVID-19 deaths with long term exposure of $PM_{2.5}$ is consistent with the previous works[20,21]. Note that, long-term $PM_{2.5}$ exposure is substantially associated with some of the comorbidities (e.g. chronic lung disease, cardiovascular disease, etc.) that may lead to COVID-19 deaths[22,23]. Similarly, a positive association between COVID-19-related deaths and $Tot\_population$ is also observed in other studies[6,24]. However, the reverse association is found for these two variables ($PM_{2.5}$ and $Tot\_population$) in the other parts of India. The explanatory variable HH_With_TCMC is found to be an important factor that may be a measure of the number of households with the upper class and rich people. A strong positive relationship is observed between HH_With_TCMC and COVID-19 death in the hotspots of eastern and western parts of India (Kolkata, North 24 Parganas, Pune, Thane, Surat, Nagpur, etc.). Note that, in those hotspots, the value of HH_With_TCMC is substantially high. An interesting observation reveals that a strong negative relationship exists between COVID-19 death and $Higher\_Edu$ in the eastern, central, and southern parts of India. It is expected that the higher educated people are well aware of the symptoms and the complications of COVID-19 that may lead to the fewer number of fatalities in those regions. Now, in some regions of the south-eastern part of India, the number of COVID-19 deaths is also seen to be high. In those regions, significant positive relationships are found between $COVID-19$ deaths and $Tot\_population$, whereas significant negative relationships are observed for the variable $Higher\_Edu$.

This research work inherits certain shortcomings that need to be resolved in future research. For instance, there may have high possibilities of under-reporting in COVID-19 death counts that may introduce bias in the study[25]. Moreover, due to data unavailability, we were not able to include some significant district-level driving factors in our study, such as health care system quality, number of hospital beds, household income, and poverty data. Despite the above-mentioned shortcomings, this is the first study that explores geographically varying relationships of COVID-19 mortality with different socioeconomic, demographic, and environmental pollution related factors in India. This research work also highlights the significance of the geographically weighted regression in the geographical analysis of the health outcome of COVID-19 disease.

Conclusion

COVID-19 pandemic is one of the most serious global public health catastrophe of the century. In this work, the geographically varying relationships between COVID-19 deaths and different potential driving factors are assessed across India. The geographical distribution of reported COVID-19 death cases is found to be heterogeneous over India. This heterogeneity in distribution is related to many underlying factors, including demographic, socioeconomic, and environmental pollution related variations between different parts of India. The GWR model makes it possible for the regression coefficients to differ across the geospace, creating geographical patterns about the strength of the relationship. The geographical heterogeneity and non-stationary of the relationships between COVID-19 deaths and the driving factors are demonstrated by mapping the local parameter estimates. The local parameter estimates reflect the quality of local model fitting and the nature of the association. The local method (GWR) yields better performance with smaller AICc as compared to the global method (OLS).

It should be noted that the impacts of other influencing factors (e.g. Meteorological factors) are not included in this work. This might be the direction for future studies.

References

1. WHO Coronavirus Disease (COVID-19) Dashboard, (2020), [Online], Available: https://covid19.who.int/.
2. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *The lancet* (2020).
3. Garg, S. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—covid-net, 14 states, march 1–30, 2020. *MMWR. Morb. mortality weekly report* **69** (2020).
4. Li, K. *et al.* The clinical and chest ct features associated with severe and critical covid-19 pneumonia. *Investig. radiology* (2020).
5. Ehlert, A. The socioeconomic determinants of covid-19: A spatial analysis of german county level data. *medRxiv* (2020).

6. Sannigrahi, S., Pilla, F., Basu, B., Basu, A. S. & Molter, A. Examining the association between socio-demographic composition and covid-19 fatalities in the european region using spatial regression approach. *Sustain. cities society* **62**, 102418 (2020).

7. Gupta, A., Banerjee, S. & Das, S. Significance of geographical factors to the covid-19 outbreak in india. *Model. earth systems environment* 1–9 (2020).

8. Sun, F., Matthews, S. A., Yang, T.-C. & Hu, M.-H. A spatial analysis of the covid-19 period prevalence in us counties through june 28, 2020: where geography matters? *Annals epidemiology* (2020).

9. Hutcheson, G. D. Ordinary least-squares regression. *L. Moutinho GD Hutcheson, The SAGE dictionary quantitative management research* 224–228 (2011).

10. Brunsdon, C., Fotheringham, S. & Charlton, M. Geographically weighted regression: A method for exploring spatial nonstationarity. *Encycl. Geogr. Inf. Sci.* 558 (2008).

11. Cressie, N. *Statistics for spatial data* (John Wiley & Sons, 2015).

12. Wheeler, D. C. & Páez, A. Geographically weighted regression. In *Handbook of applied spatial analysis*, 461–486 (Springer, 2010).

13. Fotheringham, A. S., Brunsdon, C. & Charlton, M. *Geographically weighted regression: the analysis of spatially varying relationships* (John Wiley & Sons, 2003).

14. Anselin, L. Local indicators of spatial association—lisa. *Geogr. analysis* **27**, 93–115 (1995).

15. Central Pollution Control Board, Ministry of Environment, Forest and Climate Change, Govt. of India, (2020), [Online], Available: https://cpcb.nic.in/.

16. Menard, S. *Applied logistic regression analysis*, vol. 106 (Sage, 2002).

17. Golden, B. L. & Wasil, E. A. Optimisation [by dm greig (london: Longman, 1980, 179 pp.)]. *IEEE Transactions on Syst. Man, Cybern.* **12**, 684–684 (1982).

18. Yang, W. *An extension of geographically weighted regression with flexible bandwidths*. Ph.D. thesis, University of St Andrews (2014).

19. Oshan, T. M., Li, Z., Kang, W., Wolf, L. J. & Fotheringham, A. S. mgwr: A python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS Int. J. Geo-Information* **8**, 269 (2019).

20. Wu, X., Nethery, R. C., Sabath, B. M., Braun, D. & Dominici, F. Exposure to air pollution and covid-19 mortality in the united states. *medRxiv* (2020).

21. Magazzino, C., Mele, M. & Schneider, N. The relationship between air pollution and covid-19-related deaths: an application to three french cities. *Appl. Energy* 115835 (2020).

22. Brook, R. D. *et al.* Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the american heart association. *Circulation* **121**, 2331–2378 (2010).

23. Sanyaolu, A. *et al.* Comorbidity and its impact on patients with covid-19. *SN comprehensive clinical medicine* 1–8 (2020).

24. Su, D. *et al.* Influence of socio-ecological factors on covid-19 risk: a cross-sectional study based on 178 countries/regions worldwide. *Reg. Worldw. (4/17/2020)* (2020).

25. Chatterjee, P. Is india missing covid-19 deaths? *The Lancet* **396**, 657 (2020).

Acknowledgement

This research work is supported by the project entitled- Participatory and Realtime Pollution Monitoring System For Smart City, funded by Higher Education, Science & Technology and Biotechnology, Department of Science & Technology, Government of West Bengal, India.

Conflict of Interest

The authors declare that they have no conflict of interest.

Author contributions

S.R. proposed the research topic, provided conceptual and technical guidance. A.I.M. designed the research plan, wrote the manuscript, collected the data, performed the statistical analysis. A.I.M. and S.R. both involved in the revision of the manuscript and interpretation of the results. Both the authors read and approved the final manuscript.

Figures

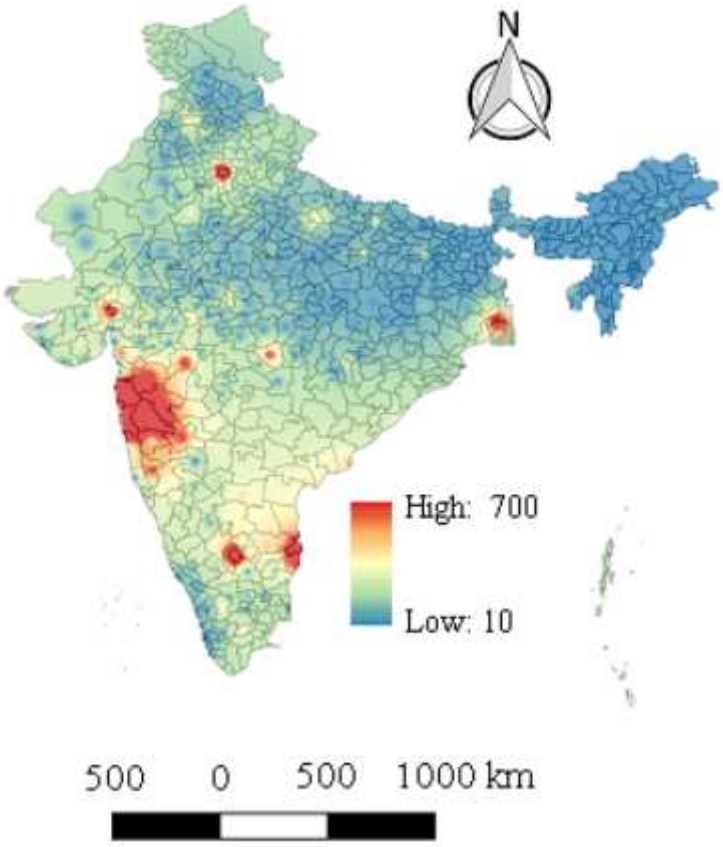

Figure 1

Geographical distribution of COVID-19 deaths across India. The spatially continuous distribution map is generated in QGIS (https://qgis.org/en/site/) by using Inverse Distance Weighting (IDW) interpolation.

Figure 2

Stepwise variable selection for geographically weighted regression (GWR).

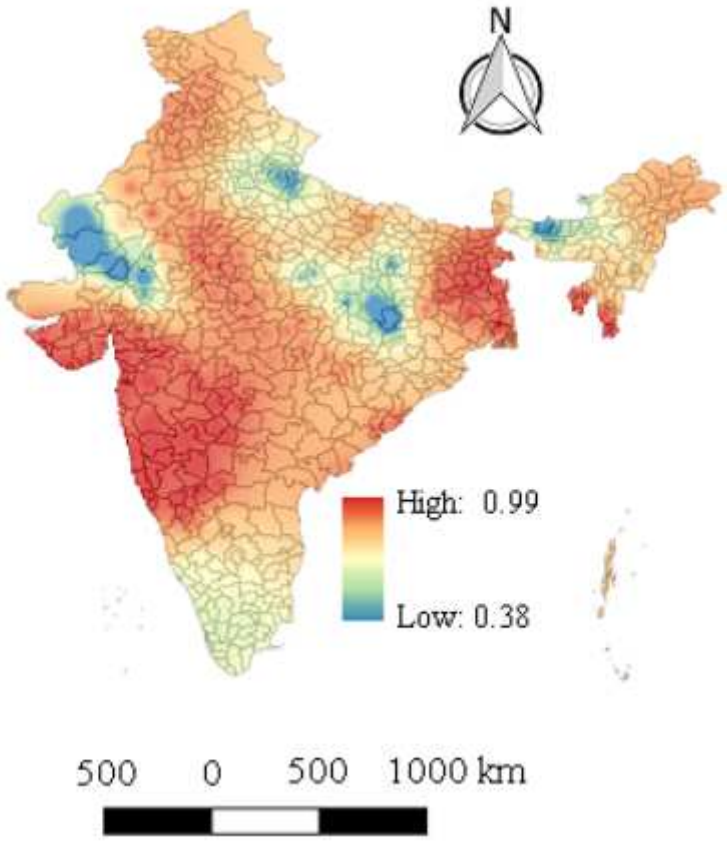

Figure 3

Geographical distribution of R 2 values for geographically weighted regression (GWR) model.

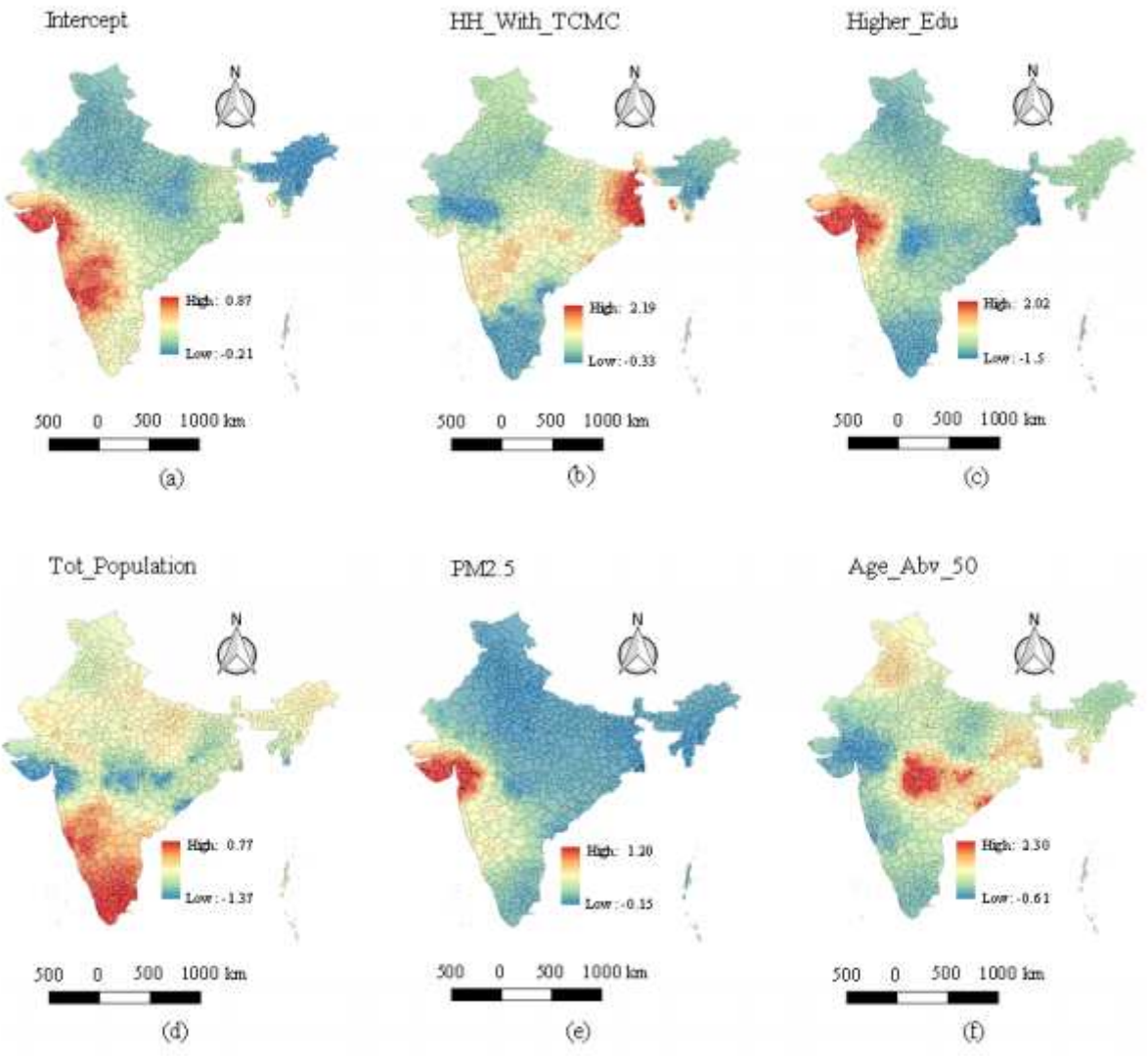

Figure 4

Local parameter estimates of geographically weighted regression (a) Intercept (b) HH With TCMC (c) Higher Edu (d) Tot population (e) PM2.5 and (f) Age Abv 50.