


# The Spatial Association of Demographic and Population Health Characteristics with COVID-19 Prevalence Across Districts in India

Sarbeswar Praharaj<sup>1</sup> , Harsimran Kaur<sup>2</sup>, Elizabeth Wentz<sup>1</sup>

<sup>1</sup>Knowledge Exchange for Resilience, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, Arizona, USA, <sup>2</sup>Department of Architecture, Planning and Design, Indian Institute of Technology (BHU), Varanasi, Uttar Pradesh, India

*In less-developed countries, the lack of granular data limits the researcher's ability to study the spatial interaction of different factors on the COVID-19 pandemic. This study designs a novel database to examine the spatial effects of demographic and population health factors on COVID-19 prevalence across 640 districts in India. The goal is to provide a robust understanding of how spatial associations and the interconnections between places influence disease spread. In addition to the linear Ordinary Least Square regression model, three spatial regression models – Spatial Lag Model, Spatial Error Model, and Geographically Weighted Regression are employed to study and compare the variables explanatory power in shaping geographic variations in the COVID-19 prevalence. We found that the local GWR model is more robust and effective at predicting spatial relationships. The findings indicate that among the demographic factors, a high share of the population living in slums is positively associated with a higher incidence of COVID-19 across districts. The spatial variations in COVID-19 deaths were explained by obesity and high blood sugar, indicating a strong association between pre-existing health conditions and COVID-19 fatalities. The study brings forth the critical factors that expose the poor and vulnerable populations to severe public health risks and highlight the application of geographical analysis vis-a-vis spatial regression models to help explain those associations.*

## Introduction

The Coronavirus pandemic (COVID-19) initially detected in December 2019 in Wuhan, China, has become a global health concern of paramount importance, affecting over 400 million cases and 5.7 million deaths as of January 10, 2022 (WHO, 2021). The first wave of the pandemic resulted in rapid virus spread in parts of Europe and North America (Desjardins, Hohl, and Delmelle, 2020). The second wave of infections originated from India fueled by a more-infectious “delta variant” of SARS CoV-2, also called B.1.617, which turned the world's second-most populous country

Correspondence: Sarbeswar Praharaj, Knowledge Exchange for Resilience, School of Geographical Sciences and Urban Planning, Arizona State University, 975 S. Myrtle Ave., Tempe, AZ 85282, USA  
e-mail: s.praharaj@asu.edu

Submitted: October 13, 2021. Revised version accepted: May 5, 2022.

into the global pandemic hotspot. According to the WHO Coronavirus Dashboard, India added over 20 million COVID-19 cases between May to June 2021, significantly more than any other country, putting immense pressure on the country's public health infrastructure. However, the official numbers only capture the limited extent of the virus spread due to the low capacity to perform large-scale testing (Subramanian et al., 2020; Zhang, Kim, and Subramanian, 2020). As of July 25, 2021, India performed 328,053 tests per million inhabitants, nearly 10 times less than that of the UK, placing it at number 106 in most testing across the global regions (Worldometer, 2021). This suggests that most cases in India are likely to go undetected. Indeed, recent community testing conducted by the Indian Council of Medical Research (ICMR) shows that the actual infection rate in the COVID-19 containment zones in the most affected cities of Mumbai, Pune, Delhi, Ahmedabad, and Indore, could be 100–200 times higher than the officially reported cases (Murhekar et al., 2021a). The ICMR studies further reveal that even in rural districts such as Ganjam and Vizianagaram, close to 40% tested positive for antibodies, indicating the infection's pervasive nature and the severity of the pandemic facing the country of 1.4 billion people. These statistics are staggering and raise significant questions around how diseases are transmitted due to geography and how people move in space.

This study investigates how spatial econometrics modeling approaches (Chi and Zhu, 2019; Sannigrahi et al., 2020; Sun et al., 2020; Andersen et al., 2021) can inform understanding of the spatial structures and the interconnections between places to shed light on the COVID-19 transmission pattern in India. We assess the impact of demographic characteristics and population health conditions on the spread of COVID-19 cases and deaths across 640 districts in India. Our goal to reveal “new geography of health” emerging from the surging COVID-19 pandemic acknowledges and reinforces the long-standing medical geography emphasis on people and places as an essential dimension of understanding and containing diseases (Glass, 2000; Praharaj and Han, 2021). Previous research found conclusive evidence that demographic factors play a vital role in determining COVID-19 cases and fatalities in the USA (Wang et al., 2020; Andersen et al., 2021; DuPre et al., 2021) and Europe (Sannigrahi et al., 2020; Perone, 2021). Being elderly is the most recognized and highest risk factor for complications from COVID-19, established through various research (Dowd et al., 2020; Subramanian et al., 2020). Preliminary data from India on COVID-19 fatalities by May 20, 2020 showed that 52% of deaths were associated with people above 60 (Joe et al., 2020). High population density also plays a detrimental role in rapid virus spread, especially in crowded urban areas where social distancing is not always possible (Hamidi, Ewing, and Sabouri, 2020; Bhadra, Mukherjee, and Sarkar, 2021). People living in slums are more vulnerable due to high density, poor access to health care services, and poverty in the global south cities (Praharaj and Han, 2019). Sannigrahi et al. (2020) found income inequality and poverty as significant influencers in geographical variations in COVID-19 fatalities. It was no surprise that more than half of the slum dwellers contracted the virus in the first two months of the pandemic in Mumbai, where population density in the Dharavi slum is more than 10 times that of Manhattan (Biswas, 2020). Access to water supply within premises is also statistically significant, as Subramanian et al. (2020) found that poor access to clean handwashing facilities is a major risk factor among communities in India. Recent literature also demonstrates that people with pre-existing health conditions and chronic illnesses are much more susceptible to the novel Coronavirus (Popovich, Singhvi, and Conlen, 2020; Zhou et al., 2020; Zhang et al., 2021).

Classical epidemiology identifies “geographical distribution” as a vital component of public health research (Pfeiffer et al., 2008). Advances in Geographical Information Systems (GIS) and Geostatistics in the last few decades revolutionized the ability to handle complex

geocoded health data with substantially improved efficiency in processing and analyzing multiple variables across spatial scales. These advancements provide medical geographers and public health authorities with new instruments to incorporate space and place in epidemiological studies to guide the identification of pandemic hotspots/clusters and the precise identification of critical vulnerability factors (Franch-Pardo et al., 2020). Many researchers adopted spatial analysis and regression-based approaches, especially in the European and the North American context, to determine geographical disparities of the COVID-19 spread while considering the spatial dependencies among various demographic, social, and economic factors (Franch-Pardo et al., 2021). Sun et al. (2020) developed a spatial analysis of the United States counties to understand how racial/ethnic composition, unemployment, income, severe housing problems, old age population, and life expectancy correlate with the COVID-19 period prevalence. They found that nonwhite-white population segregation, life expectancy, and population density are positively associated with the COVID-19 prevalence. Sannigrahi et al. (2020) found a strong association between demographic factors (income and poverty) and COVID-19 fatalities in the European region using a spatial regression approach. Similarly, Mollalo, Vahedi, and Rivera (2020) examined the relationship between 35 environmental, socioeconomic, and demographic variables and COVID-19 incidence through GIS-based spatial modeling to find that ethnicity and income disparities positively control the geographic distribution of COVID-19 cases in the United States. Kapitsinis (2020), while studying 119 regions across nine Western European countries, argues that “geography” plays a vital role in the viral spread and that variations in air quality, demography, global interconnectivity, urbanization trends, and health expenditure all contributed to the regionally uneven mortality rate.

Previous research has highlighted the challenges involved in implementing these spatial analysis models in low and middle-income countries (LMICs), where representative and granular data on established COVID-19 risk variables are severely lacking (Subramanian et al., 2020; Bhadra, Mukherjee, and Sarkar, 2021; Praharaj et al., 2020). Interestingly, many LMICs, such as India, are severely impacted by the deadly second wave of the pandemic, with the World Bank estimating that the lockdowns and economic disruptions might push up to 50 million people into extreme poverty in South Asia (World Bank, 2020). Greater epidemiologic intelligence is required through micro-level geographical analysis to support decision-makers in these regions in identifying high-risk areas and communities to strategize a timely and cohesive response to the evolving crisis. Subramanian et al. (2020) published a study demonstrating a geo-visualization approach for identifying areas with potentially differential susceptibilities to COVID-19 using data from the National Family Health Survey (NFHS) in India, which was made available through the Demographic and Health Surveys (DHS) Program (Corsi et al., 2012). In almost 90 countries, the DHS provides accurate and standardized data on population health. However, no additional research has been conducted to establish how non-clinical risk factors interact with demographic characteristics, environmental conditions, and access to infrastructure in India’s districts or other LMICs to determine how geographical disparities influence COVID-19 outcomes. This article builds on this research gap by examining the spatial association between multiple factors of demographics, and pre-existing population health determinants with COVID-19 cases and deaths across districts in India. The study’s outcomes can potentially be used to develop prevention strategies and augment public health capacities in critical need geographies where health systems cannot be substantially enhanced in time to deal with a surging pandemic and its associated economic loss.

The goal of this study is to analyze the interaction of the variables over time and geography in determining the COVID-19 outcomes in India. We join geographers and public health scholars to call for attention to the role of place-based characteristics in the disease spread and the spatial relationships or interconnections between places to build a comprehensive knowledge of the potential determinants of the novel disease.

## Materials and methods

For this study, we assembled a district-level data set for 640 districts in India. We consistently used the list of districts as reported in Census 2011 administered by the Office of the Registrar General and Census Commissioner, India (ORGI, 2020). We used Geo-referenced GIS district boundary layers available through the National Remote Sensing Centre of ISRO for performing spatial analysis. The following subsections outline the dependent and independent variables used in the research, the rationale for their selection, and the spatial regression approach for modeling the data.

### Dependent variables

The dependent variable is the COVID-19 period prevalence (the number of cumulative confirmed cases and deaths per 100,000 population) in a district as of May 9, 2021. We obtained the district-level counts of COVID-19 cases and deaths from COVID-19 India Tracker (COVID-19 India Org Data Operations Group, 2020), a crowd-sourced database for real-time COVID-19 statistics and patient tracing in India. The dataset made available through an open (CC-BY-4.0) license is the most comprehensive source available for district-level COVID-19 information in India. It combines the daily data from the Indian Ministry of Health and Family Welfare and data published through the state health department bulletin from all the 36 states and union territories in India. The other sources providing daily data, such as the John Hopkins University's COVID-19 dashboard (see <https://coronavirus.jhu.edu/map.html>) or the WHO Coronavirus (COVID-19) Dashboard (see <https://covid19.who.int/>) only provide data up to state level in India.

### Independent variables

Initially, we identified 16 independent variables across demographic, environmental, and population health characteristics for regression modeling. Table S1 summarizes these variables that have gone through a series of statistical analyses, including computation of variable inflation factors (VIFs), principal component correlation matrix, and a stepwise forward regression approach to identify multicollinearity and redundancy in the dataset. Through these processes, a total of six variables were chosen that were generally uncorrelated and followed a normal distribution. Those are population density in persons/km<sup>2</sup>, percentages of the elderly population (>65 years), percentages of the population living in slums, percentages of households without access to water supply, percentages of women who are overweight or obese, and percentages of women with high blood sugar.

We used the 2011 District Census Handbook (DCH) to collect demographic indicator data for the 640 districts selected in this study. The DCH provides decadal census data (ORGI, 2020) in a structured and accessible format across the states and districts in India, with the latest information available for 2011. Indicators including population density, share of population over 65 years, population living in slums, and households without access to water supply were collected from this source.

We leveraged district-level information on population health characteristics from the fourth NFHS Survey of 2015–2016. Two population health indicators were chosen for this analysis. These are the percentages of women who are overweight or obese (Body Mass Index greater than or equal to 25.0) and the percentages of women with high blood sugar (glucose level  $\geq 126$  mg/dL after 8 h of opportunistic fasting). The NFHS follows a fairly robust sampling strategy and data collection procedure outlined through the DHS program implemented in over 90 LMICs. We applied to the DHS to access this data by registering our project through <https://dhsprogram.com/> and were provided secured links for free downloading of the data for use in this specific project. A limitation of the population health data used in this study is that it represents district-level population health characteristics for women. Data produced through NFHS provides representative data for men for only up to state-level (IIPS and ICF International, 2017), and our focus was on district-level estimates.

### Spatial regression modeling approach

We implement a spatial regression model for estimation and diagnostic checking of the spatial effects, including spatial dependence, spatial autocorrelation, and spatial heterogeneity (Fotheringham and Rogerson, 2012). Spatial regression approaches are highly effective and reliable when variables are locally varying, spatially dependent, and autocorrelated, making them a definite choice for spatial epidemiological studies where we find high geographical dependencies and correlations between observations (Chi and Zhu, 2019; Sannigrahi et al., 2020; Sun et al., 2020; Andersen et al., 2021). Researchers prefer spatial regression over traditional statistical functions such as factor analysis and principal component analysis (Praharaj, Han, and Hawken, 2018). The latter assumes that samples used in these models are independent of one another. The spatial regression models effectively estimate the influence of independent factors on target variables by differentiating the spatial dependence while including the lag and error components of independent features.

This study analyses the data through the linear Ordinary Least Square (OLS) regression model and three spatial regression models – Spatial Error Model (SEM), Spatial Lag Model (SLM), and Geographically Weighted Regression (GWR). These models help us examine how the various factors of demographic, environmental, and population health characteristics shape the pattern of COVID-19 cases and deaths across Indian districts. While the SEM, SLM, and OLS models support the measuring of global interaction between the selected variables and the predictor, GWR provides a more unbiased assessment of local associations (Fotheringham, 2012).

The OLS regression models examine the (non) spatial relationships between the set of control and response variables assuming homogeneity and spatial nonvariability across the study area under investigation (Mollalo, Vahedi, and Rivera, 2020; Sun et al., 2020).

$$y_i = \beta_o + x_i\beta + \varepsilon_i$$

where  $i$  and  $y_i$  are the COVID-19 incidence parameters (cases and deaths),  $\beta_o$  is the intercept,  $x_i$  is the vector of selected demographic and population health variables,  $\beta$  is the vector of regression coefficients, and  $\varepsilon_i$  is a random error. The fundamental function of OLS is to optimize the regression coefficients ( $\beta$ ) by reducing the sum of squared prediction errors. We deployed OLS first, to check the spatial autocorrelation of the model's residuals to determine the appropriateness of a spatial regression model.

The SLM, in comparison, incorporates spatial dependency between the parameters into the regression model that help us examine how the infection burden in a district is influenced by the infection burden in adjacent districts. SLM models, by default, assume a close association between the response and control variables.

$$y_i = \beta_o + x_i\beta + \rho W_i y_i + \varepsilon_i$$

where  $\rho$  is the spatial lag parameter, and  $W_i$  is a vector of spatial weights (a row of the spatial weights matrix). The weight matrix ( $W$ ) of SLM indicates the neighbors at location  $i$  and connects one independent variable to the explanatory variables in feature space (Sannigrahi et al., 2020).

The SEM model estimates the correlations and strengths of the relationship of the OLS residuals/errors between neighboring districts, assuming that there is a spatial dependence (Andersen et al., 2021). The typical motivation for this is that unmodeled effects spill over across observation units and hence result in spatially correlated errors. Hence, the residuals of OLS are separated into two spatial components – error term and a random error term (for satisfying the assumption in the modeling). A limitation of application in our study is that the model assumes that the error term in the models has constant variance.

$$y_i = \beta x_i + u_i$$

$$u_i = \lambda w_i u_j + \varepsilon_i$$

where  $u_i$  and  $u_j$  are the error terms at locations  $i$  and  $j$ , respectively, and  $\lambda$  is the coefficient of spatial component errors.

In contrast to the global regression models (OLS, SLM, and SEM), GWR estimates the local interaction among the control and response variables after integrating the geo-referenced data layers. GWR is most effective when the data under investigation are spatially heterogeneous or nonstationary, as is the case for most social and epidemiologic studies. Fotheringham (2012) suggested that the GWR can easily compute locally varying parameter estimates, and thus found to be highly effective to produce detailed spatially explicit maps of locational variations in relationships.

$$y_i = \beta_{i0} + \sum_{j=1}^m \beta_{ij} X_{ij} + \varepsilon_i$$

$$i = 1, 2, 3 \dots \dots \dots n.$$

where  $y_i$  is the value of response variables,  $\beta_{i0}$  is the intercept,  $\beta_{ij}$  is the  $j$ th regression parameter,  $X_{ij}$  is the value of the  $j$  th explanatory variable,  $\varepsilon_i$  is the random error. Regarding the kernel selection and defining local weight matrix in the GWR model, the ‘‘Gaussian fixed kernel’’ with AICc (corrected Akaike Information Criterion) estimated bandwidth was chosen. The selection was based on previous studies that found in case–control studies of a rare disease bi-square kernel is not secure even for adaptive kernel and Gaussian adaptive kernel serves as a securer option (Nakaya et al., 2012). The neighborhood size is considered as a function of a specified number of neighbors; in this study, we have selected the neighborhood with the Golden search based on minimizing the value of the Akaike Information Criterion (AIC). For defining the global spatial weight, the first-order Queens’ contiguity approach was adopted for the other spatial regression models (OLS, SEM, and SLM).

GeoDa software was used to conduct OLS, SEM, and SLM spatial regression analyses. GeoDa was also used to perform the Global Moran’s I spatial dependence test, which was used

**Table 1.** Descriptive Statistics of Variables Used in this Study, as of May 9, 2021 ( $n = 640$ )

Variable	Mean	SD	Minimum	Maximum	VIF
Confirmed cases per 100,000 (logged)	3.00	0.50	0.00	5.02	–
Confirmed deaths per 100,000 (logged)	0.93	0.53	–0.58	3.18	–
Deaths/Cases Ratio per 100,000	0.01	0.01	0.00	0.04	–
Population density in persons/km <sup>2</sup> (logged)	2.57	0.53	0.00	4.66	1.36
% of elderly population (>65 years)	4.24	1.31	1.07	11.17	1.27
% of slum population	3.80	4.88	0.00	49.38	1.26
% of population with water supply away from premises	20.19	12.35	0.00	74.79	1.70
% Women who are overweight or obese	18.75	9.08	0.00	48.70	2.29
% Blood sugar level among Women – high	5.59	2.11	0.00	12.50	1.40

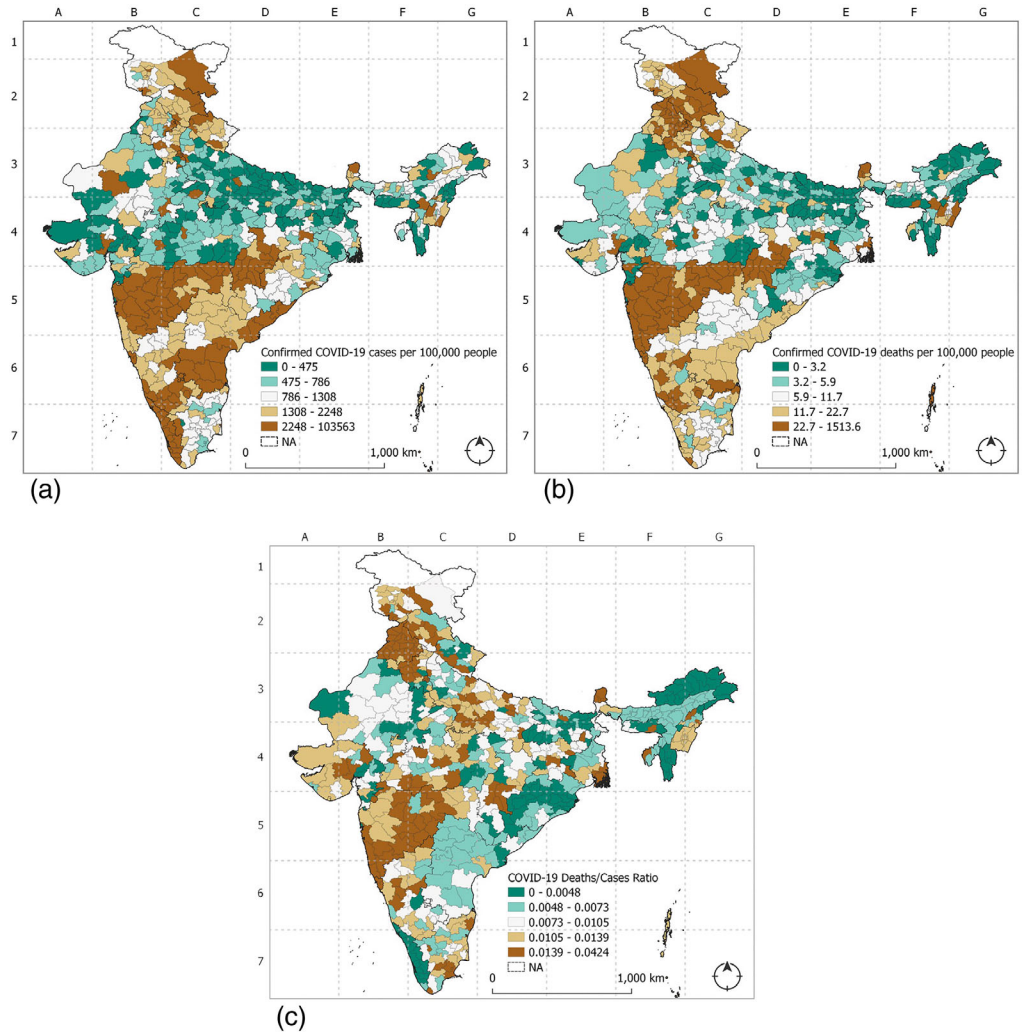
to determine the statistical significance of the distribution's geographic clusters. ArcMap 10.4.1 was used to visualize and map the data and implement the GWR regression model.

## Results

Table 1 presents the descriptive statistics of the variables, including the variance inflation factors. We emphasize that multicollinearity is not a concern in our analysis as the variance inflation factors are all less than 2.5, which is often considered a building block for robust spatial regression models. The district-level variation of COVID-19 cases and deaths (per 100,000 population) and cases to death ratio as of May 9, 2021, is presented in Fig. 1. The districts are categorized by quantiles produced through ArcMap 10.4.1. The map shows that districts with high COVID-19 cases and deaths are clustered in the western Indian state of Maharashtra and in the adjoining districts of Chhattisgarh and Goa. The districts of Mumbai, Nagpur, and Pune have the highest rates of cases and deaths in this region. A large concentration of patients is also visible among the neighboring districts of Kerala, Karnataka, and Andhra Pradesh in the southern parts of the country, with maximum incidences in Yanam, Bangalore, Kozhikode, Chennai, and Ernakulam. Considering the COVID-19 deaths in the northern stretch of India, a significant part of the state of Punjab, Haryana, Delhi, and Himachal Pradesh are severely affected. By contrast, districts with a low prevalence are concentrated in the Eastern and Central Indian states of Bihar, Madhya Pradesh, Jharkhand, and West Bengal. These heterogeneous distributions of COVID-19 cases and deaths can be linked with the “districts” underlying socio-demographic, housing and environmental, infrastructural, and population health characteristics.

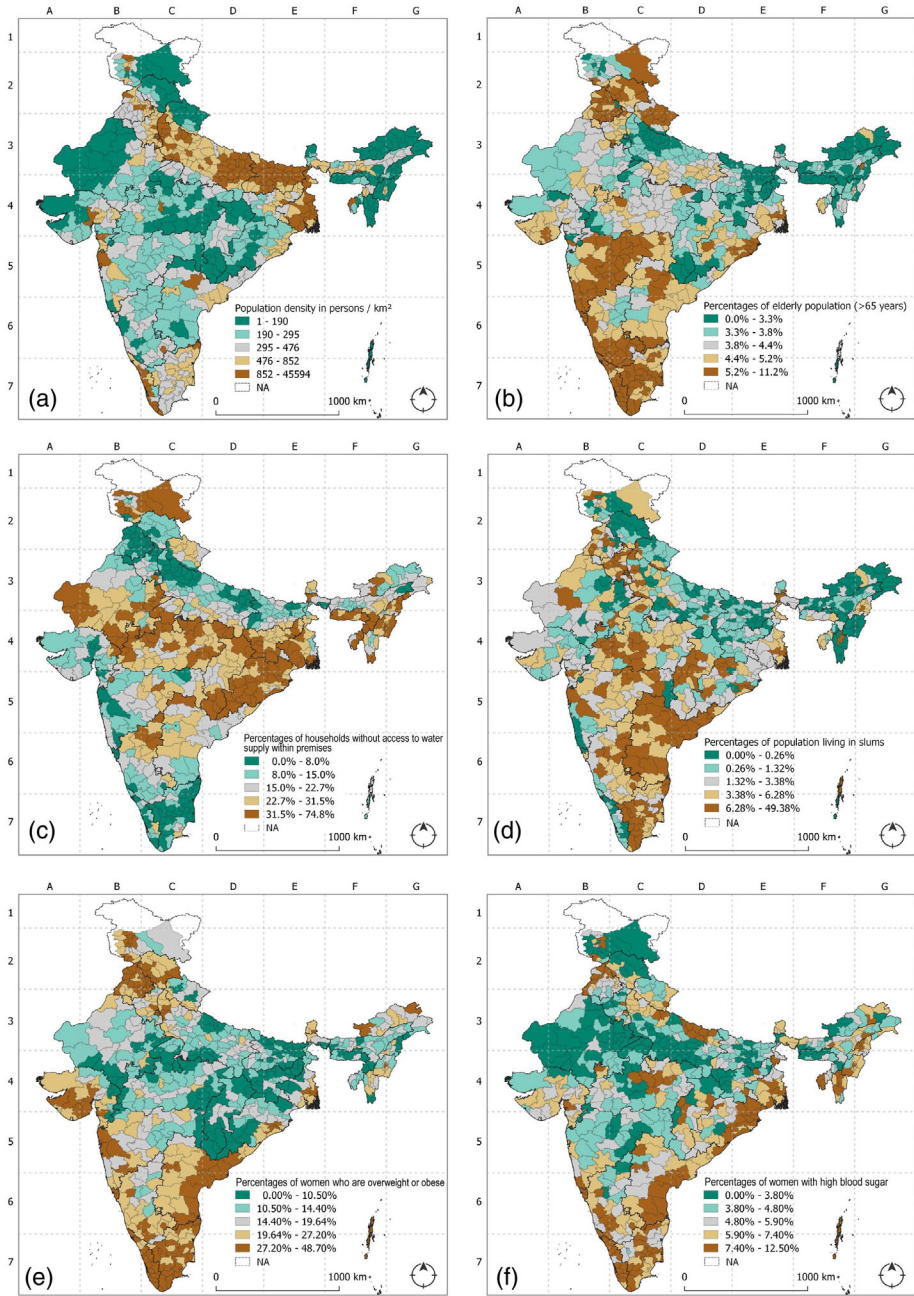
Figure 2 presents a visual analysis of the six variables entered in the regression model, providing a snapshot of district Geographies of non-clinical risk correlates. Fig. S1 identifies the name and boundaries of the districts in India, which enable a better interpretation of all the maps presented in this article. We find that the districts in Delhi had some of the highest population density, with a maximum of 36,155 persons/km<sup>2</sup>. Broadly, a very high population density is visible in Bihar, West Bengal, and Kerala, where 29 out of the 71 districts were in the top-10%. On average, 4.24% of the population across districts are older than 65 years. Maharashtra, Kerala, Karnataka, Goa, and Punjab had many districts with very high percentages of the aging

## Geographical Analysis



**Figure 1.** Spatial distribution of the COVID-19 cases per 100,000 people (a), deaths per 100,000 people (b), and death/cases ratio (c) by quantiles, as of May 9, 2021. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)].

population. Fifteen out of its 35 districts in Maharashtra and 9 of the 14 districts in Kerala are in the top 10% according to the percentage of elderly. The percentages of slum population among Indian districts are on average 3.80%, going up to a maximum of 49.38%. Andhra Pradesh, Tamil Nadu, Maharashtra, Chhattisgarh, and Madhya Pradesh show a maximum number of districts in the highest quantiles (>6.3%). Some of the largest shares of slum population are observed in Mumbai (49%), Kolkata (31%), Bhopal (30%), Hyderabad (29%), and Chennai (27%). On average, 20.19% of households across Indian districts do not have a water supply connection within premises. Fig. 2 suggests that the lack of access to the water supply is acute in several districts within Odisha, Madhya Pradesh, and Rajasthan. Twelve out of the 14 districts in Kerala and 21 of the 30 districts in Tamil Nadu fall in the highest quartiles range (>27%) on the share of women population who suffer from obesity. Ten districts in Andhra Pradesh and 5 in Maharashtra



**Figure 2.** Geospatial analysis of the independent variables. (a) population density in persons/km<sup>2</sup>, (b) percentages of elderly population above 65 years, (c) percentages of households without access to water supply within premises, (d) percentages of population living in slums, (e) percentages of women who are overweight or obese with BMI  $\geq 25.0$  kg/m<sup>2</sup>, and (f) percentages of women with high blood sugar (glucose level  $\geq 126$  mg/dL). [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)].

also fall in this category. In a similar pattern, Kerala, Goa, Tamil Nadu, and Andhra Pradesh show the maximum number of districts having the highest percentages of women (>7.5%) with high blood sugar.

The OLS and spatial modeling results are summarized in Table 2, and several findings are notable. First, population health factors, including the elderly population, obesity, and high blood sugar are significant variables in determining the COVID-19 impact in Indian districts, although the magnitude of the coefficients across models varies. Second, the models highlight that there is a positive association between COVID-19 and the concentration of slums, and households unserved with water supply. These associations are consistent across all models and robust to the specification of spatial dependence. Results from this research align with previous studies conducted in the United States, where Sun et al. (2020) found that factors such as county-level racial/ethnic composition, life expectancy, and population density are positively associated with the COVID-19 prevalence, and the spatial effects were consistent across models.

We map the  $R^2$  values to discover the association and effect of the independent variables with the COVID-19 cases (Fig. 3) and deaths (Fig. 4) for each of the 640 districts under investigation. The largest proportion of the variation considering local  $R^2$  values for the cases was explained by the percentages of the population living in slums in Thoothukudi ( $R^2 = 0.660$ ), Tirunelveli ( $R^2 = 0.657$ ), and Krishnanagar ( $R^2 = 0.647$ ). We find a high explanatory power of this variable in parts of Maharashtra, Kerala, Madhya Pradesh, West Bengal, and Chhattisgarh. Percentages of households without access to water supply within premises had substantial predictability for districts such as Rewari ( $R^2 = 0.760$ ), Jhajjar ( $R^2 = 0.754$ ), Rohtak ( $R^2 = 0.736$ ), and large parts of the state of Odisha, Karnataka, Chhattisgarh, and Madhya Pradesh. A strong interaction effect between percentages of women with high blood sugar and the COVID-19 cases was observed in Nicobar ( $R^2 = 0.422$ ), Krishnanagar ( $R^2 = 0.303$ ), and Surguja ( $R^2 = 0.2938$ ). Blood sugar substantially impacts COVID-19 case prediction predominantly in southern India and districts surrounding the national capital region of Delhi. Obesity and the COVID-19 cases were very well linked in Krishnanagar ( $R^2 = 0.670$ ), North 24 Parganas ( $R^2 = 0.658$ ), Kolkata ( $R^2 = 0.621$ ), and several districts in West Bengal and Madhya Pradesh.

The highest association between COVID-19 death and obesity was accounted for several districts of Karnataka, West Bengal, Uttar Pradesh, Madhya Pradesh, and Andhra Pradesh, with Nicobar ( $R^2 = 0.791$ ), Bahadurganj, ( $R^2 = 0.776$ ), and Darjeeling ( $R^2 = 0.765$ ) leading the group. Percentages of the population living in slums could explain a sizeable proportion of the variation in the dependent variable (COVID-19 deaths) in the districts of Jodhpur ( $R^2 = 0.721$ ), Tonk ( $R^2 = 0.717$ ), Vidisha ( $R^2 = 0.709$ ), and many others in Rajasthan and Madhya Pradesh. An emerging association between percentages of households without access to water supply within premises and the COVID-19 death was observed in Mahasamund ( $R^2 = 0.725$ ), Kolhapur ( $R^2 = 0.724$ ), and Solapur ( $R^2 = 0.718$ ), and in the adjoining cluster of districts in the state of Odisha, Andhra Pradesh, Madhya Pradesh, and Chhattisgarh. The sizeable impact of the percentages of the elderly population (>65 years) was observed on the dependent variable in Uttar Dinajpur ( $R^2 = 0.735$ ), Araria ( $R^2 = 0.702$ ), Bahadurganj ( $R^2 = 0.699$ ), and the majority districts in the states of Maharashtra and Goa.

Table 3 summarizes and compares the spatial regression estimates across the global (OLS, SLM, and SEM) and local (GWR) models. The GWR model emerged as the strongest of the four, with a significant increase in predictive power and model fit with the data. For the COVID-19 case factor, the prediction accuracy in the GWR model goes up to 45% from just 11.4% in the OLS model. The six variables predicted just over 10% of the model variance on

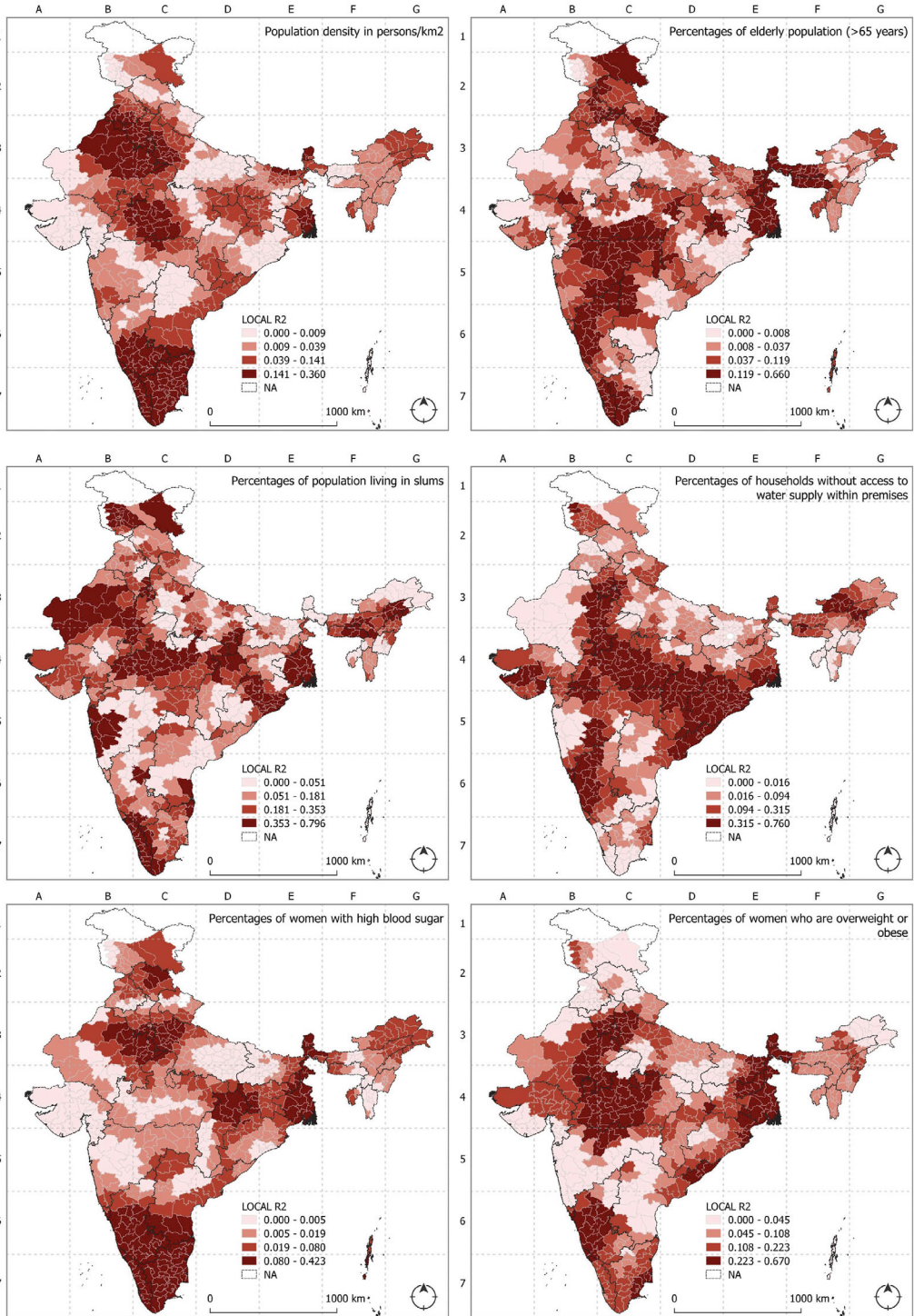
**Table 2.** OLS, Spatial lag, Spatial error, and GWR Model for the COVID-19 Cases (Logged), Deaths (Logged), And Death/case Ratio, as of May 9, 2021

Factor	Variable	OLS		Spatial lag model		Spatial Error model		GWR model	
		Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Cases	Confirmed cases per 100,000 (logged)	-	-	0.134***	0.039	-	-	-	-
	CONSTANT	2.407***	0.143	2.118***	0.166	2.557***	0.171	-	-
	Population density in persons/km <sup>2</sup> (logged)	0.002	0.039	-0.004	0.039	0.043	0.048	0.160	0.404
	% of elderly population (>65 years)	0.088***	0.015	0.073***	0.016	0.062***	0.018	0.058	0.381
	% of slum population	0.006	0.004	0.005	0.004	0.000	0.004*	0.004*	0.288
	% of population with water supply away from premises	-0.001	0.002	-0.002	0.002	-0.006***	0.002*	-0.012**	0.346
	% Women who are overweight or obese	0.015***	0.003	0.014***	0.003	0.013***	0.003*	0.022*	0.376
	% Blood sugar level among Women - high	-0.012	0.010	-0.008***	0.010	-0.006	0.011	0.025	0.414
	$\rho$ (spatial lag parameter)			0.133731					
	$\lambda$ (spatial error parameter)					0.422841***	0.0492292		
Deaths	Confirmed deaths per 100,000 (logged)	-	-	0.479***	0.040	-	-	-	-
	CONSTANT	0.106	0.135	0.036	0.118	0.293*	0.162	-	-
	Population density in persons/km <sup>2</sup> (logged)	0.035	0.037	0.016	0.032	0.075*	0.046	0.208	0.376
	% of elderly population (>65 years)	0.107***	0.015	0.052***	0.013	0.048***	0.017	0.085	0.327

Table 2. Continued

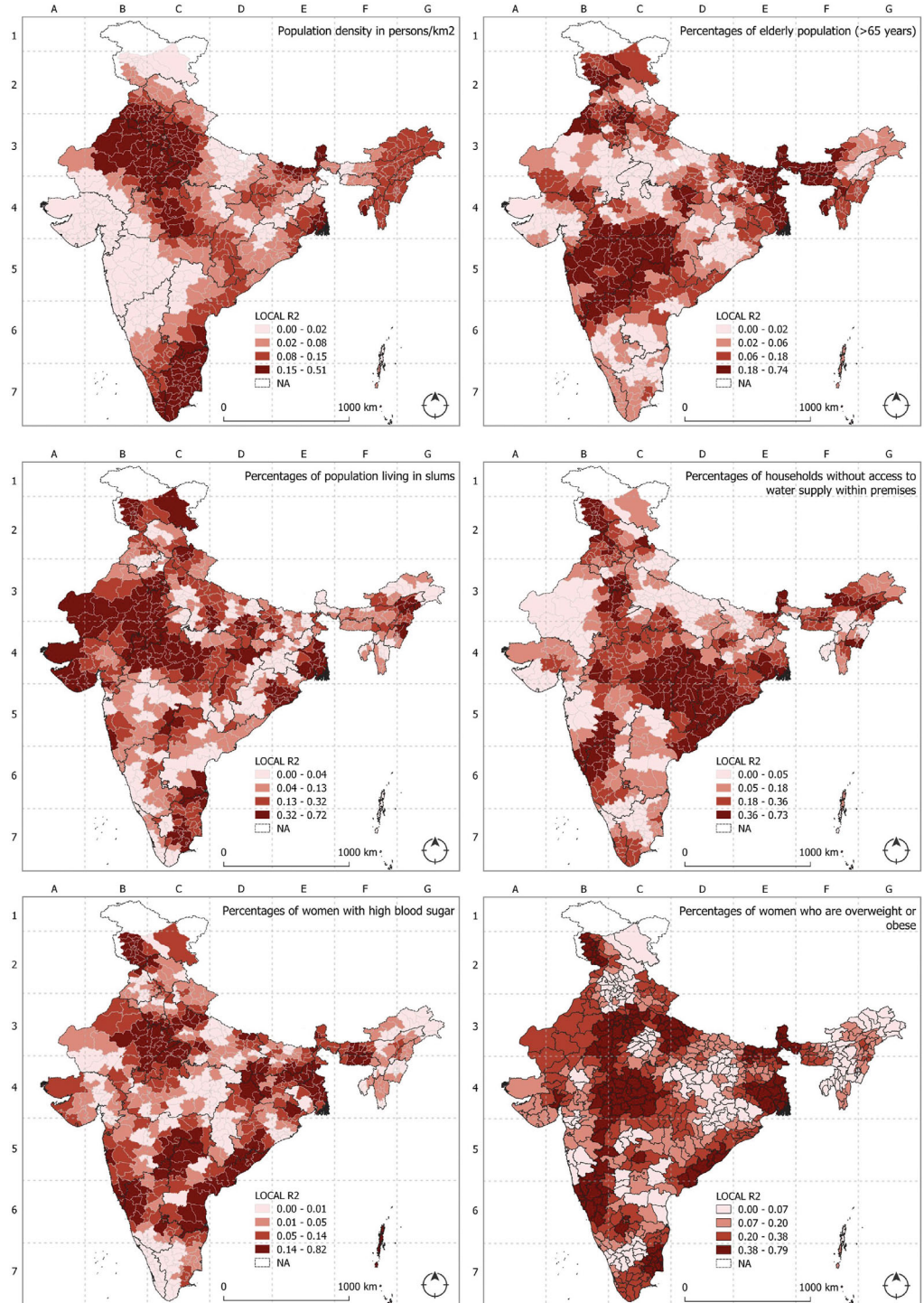
Factor	Variable	OLS		Spatial lag model		Spatial Error model		GWR model	
		Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
	% of slum population	0.018***	0.004	0.012***	0.003	0.006*	0.004	0.026	0.288
	% of population with water supply away from premises	0.000	0.002	-0.002	0.002	-0.005***	0.002*	-0.013*	0.3
	% Women who are overweight or obese	0.023***	0.003	0.015***	0.003	0.021***	0.003	0.031	0.296
	% Blood sugar level among Women - high	-0.038***	0.009	-0.017**	0.008	-0.011	0.010	0.035	0.332
	$\rho$ (spatial lag parameter)			0.479					
	$\lambda$ (spatial error parameter)					0.629***	0.038		
Deaths/Cases	Deaths/Cases ratio	-	-	0.6035***	0.0379	-	-	-	-
	CONSTANT	0.0040**	0.0018	0.0009	0.00146	0.0041	0.0021	-	-
	Population density in persons/km <sup>2</sup> (logged)	0.0007	0.0005	0.0003	0.00040	0.0007	0.0006	0.0008	0.0052
	% of elderly population (>65 years)	0.0008***	0.0002	0.0002	0.00016	0.0002	0.0002	0.0012	0.0049
	% of slum population	0.0000	0.0001	-0.0000	0.00004	-0.0001	0.0000*	0.0001*	0.0052
	% of population with water supply away from premises	-0.0000	0.0000	-0.0000	0.00002	-0.0000	0.0000*	-0.0001*	0.0051
	% Women who are overweight or obese	0.0002***	0.0000	0.0001***	0.00003	0.0001***	0.0000*	0.0001*	0.0050
	% Blood sugar level among Women - high	-0.0004***	0.0001	-0.0000	0.00010	0.0001	0.0001	0.0003	0.0051
	$\rho$ (spatial lag parameter)			0.0635					
	$\lambda$ (spatial error parameter)					0.6503***	0.0370		

Note: \*P < 0.05, \*\*P < 0.01, \*\*\*P < 0.001.



**Figure 3.** The district-level effect of the variables on the COVID-19 cases derived from the GWR model. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)].

# Geographical Analysis



**Figure 4.** The district-level effect of the variables on the COVID-19 deaths derived from the GWR model. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)].

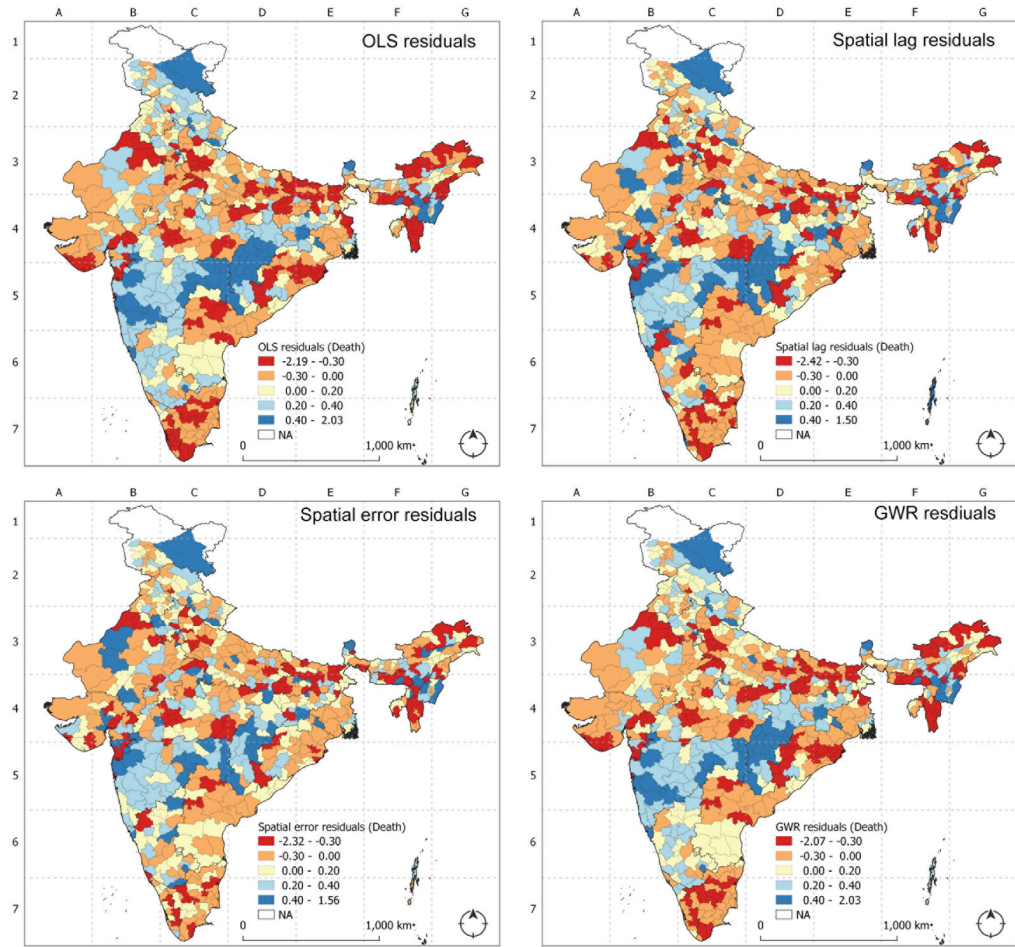
**Table 3.** Overall Summary of Spatial Regression Models Indicating the Linkages Between the Variables and Total COVID-19 Cases and Deaths Across India, as of May 9, 2021

Factor	Variable ( <i>n</i> )	Parameters	Spatial Regression Models			
		Parameters	OLS	SLM	SEM	GWR
Cases	<i>n</i> = 6	R-squared	0.114	0.125	0.234	0.452
		Adjusted R-squared	0.066	–	–	0.366
		AIC	894.226	862.147	800.928	704.369
		SIC	903.149	875.531	809.851	–
Death	<i>n</i> = 6	R-squared	0.114	0.433	0.474	0.622
		Adjusted R-squared	0.113	–	–	0.534
		AIC	916.049	690.077	656.755	591.270
		SIC	924.972	703.461	665.678	–
Deaths/ Cases	<i>n</i> = 6	R-squared	0.126	0.414	0.426	0.550
		Adjusted R-squared	0.118	–	–	0.433
		AIC	–4799.310	–4998.200	–5003.020	–5884.828
		SIC	–4768.080	–4962.510	–4971.790	–
Cases	Observed Moran's <i>I</i> for residuals <i>I</i>		0.191	0.114	–0.034	0.166
Death			0.342	0.022	–0.048	0.309
Deaths/ Cases			0.400	–0.034	–0.053	0.310
Cases		<i>P</i> -value Level of significance	0.001	0.002	0.114	0.001
Death		0.001	0.172	0.039	0.001	
Deaths/Cases		0.001	0.088	0.021	0.001	

Note: *n* = 6 includes six variables: (1) Population density in persons/km<sup>2</sup> (logged), (2) % of elderly population (>65 years), (3) % of slum population, (4) % of population with water supply away from premises, (5) % women who are overweight or obese (BMI ≥25.0 kg/m<sup>2</sup>), and (6) % blood sugar level among Women – high.

COVID-19 death in the OLS model, which rose to 62.2% in the GWR model. A similar trend was observed in the case of the death/case ratio factor, where the explanatory variables predicted the death/case ratio with 55% accuracy against the 12.6% observed in the OLS model. There is also intra-model variation in the effects among the three factors. For example, in the GWR model, 45.25% variance was explained by the six variables for COVID-19 cases, where the same measures accounted for 62.2% of the model variance for COVID-19 deaths. Overall, we find that the model variances were better captured for deaths with significantly increased explanatory capacity shown across the models.

A measure of the AIC and Schwarz criterion (SIC) is provided in Table 3 to determine the quality of each model relative to the others. The AIC and SIC scores are used as model selection criteria to compare the suitability of OLS, SLM, SEM, and GWR models. This selection criteria seek a compromise between model fit and model complexity, and these models have different degrees of complexity. The assessment highlights that the OLS model produced the highest AIC values, SLM and SEM models show moderate estimates, and the GWR consistently has the lowest AIC statistics. The lower value of AIC suggests the GWR model most efficiently meets the selection criteria. A similar trend is observed for SIC goodness of fit prediction, with OLS emerging as the weakest model. Better explanatory capacity demonstrated through higher coefficient values and lower AIC estimates reveals that the local GWR model is superior to global spatial regression models in accurately measuring spatial dependence among the observations.

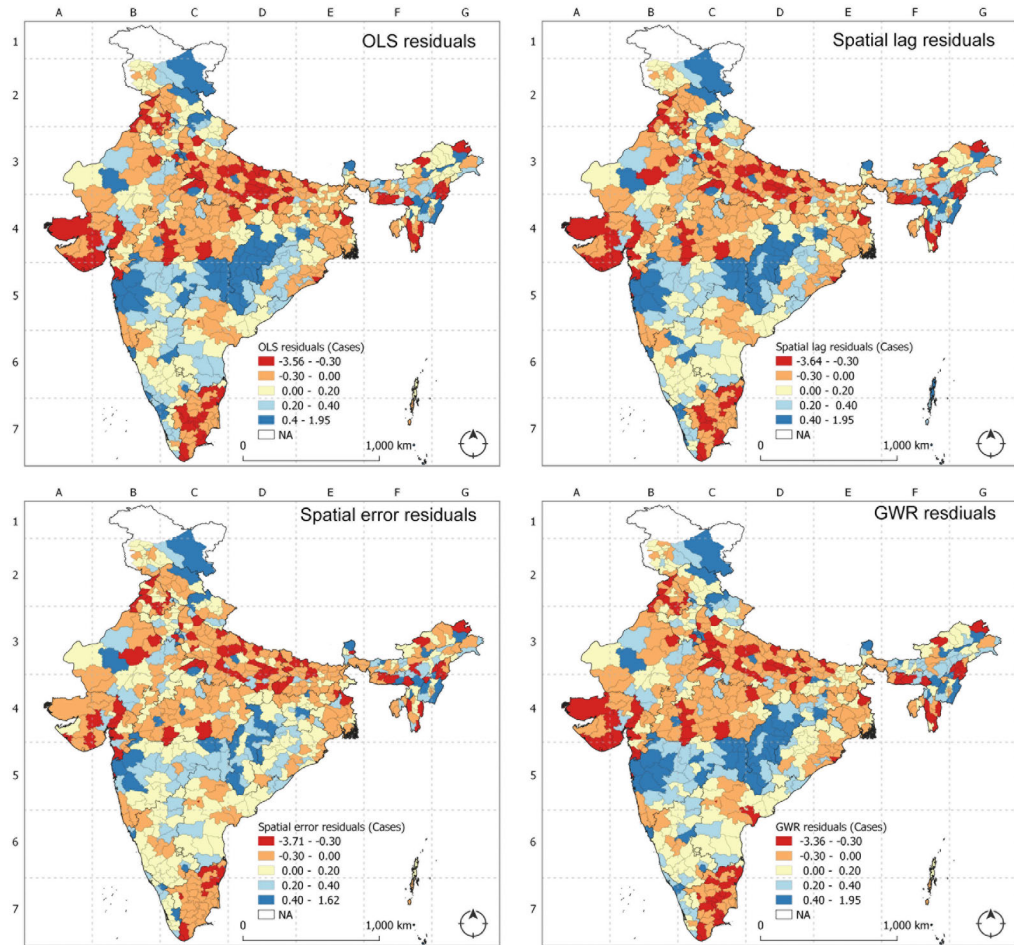


**Figure 5.** Map of residuals of OLS, spatial lag, spatial error, and GWR model with COVID-19 cases. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)].

The residuals of all four models (OLS, SLM, SEM, and GWR) are visualized in Fig. 5 for COVID-19 cases and Fig. 6 for deaths. Although we found significant variability in the  $R^2$  approximation, the spatially varying coefficient values, AIC estimates across the regression models, and the residuals from the models show similar patterns. The results show that the residual values are generally small across districts suggesting that the regression analysis has successfully explained the essential part of the variation of the dependent variable. However, the high residual values presented through the model in the Central Indian districts and a significant observed Moran's I for some models indicate that there is some unexplained difference that is not fully captured by the independent variables in the regression model and there is a need for further research to this end.

## Discussion

This study offers a novel application of the spatial regression models to analyze COVID-19 risk factors in India. While the trajectory and spatiality of COVID-19 in India are reported before



**Figure 6.** Mapping of residuals of OLS, spatial lag, spatial error, and GWR model with COVID-19 deaths. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.com)].

in scholarly literature, our study is the first to detect the stark implications of chronic health conditions on higher fatalities in India. Globally, our finding is the first to detect the interaction of a higher share of slum population on increasing infection rates in India, which have significant implications for other LMICs in South Asia, Africa, and Latin America. The spatial regression findings suggest that our model specifications reasonably explain the variations in COVID-19 cases and deaths across districts; that is, a large majority of districts have relatively small residuals in our analysis. However, it should be emphasized that the variables used in the study cannot fully account for the spatial pattern, especially in districts with very high COVID-19 outbreaks, some of which show higher residual values (marked with red color in Figs. 5 and 6). Through the map of residuals, we find that the model fits very well in one district but fits poorly in neighboring districts, indicating that they retain strong spatial dependencies. It suggests that leaving unaccounted parts of the spatial structure leaves plenty of room for improving the predicted values. Similarly, the results show districts, where the COVID-19 cases and death are significantly underpredicted, are often adjacent to districts where the model overpredicts. This variable pattern is visible across North and North-Eastern parts of India and offers a stark

contrast with the good model fit across much of the Southern and Central Indian districts. These findings support existing research; for example, Mollalo, Vahedi, and Rivera (2020) suggest that spatial heterogeneity of COVID-19 is fairly common in United States counties. While scholars previously noted the importance of spatial econometric models and autocorrelation, little research has considered the spatial lag, spatial error, and GWR simultaneously. Further, there is a lack of geographical research on how preexisting health conditions influence COVID-19 outcomes in the global south region. Our study advances the rapidly evolving literature by filling these gaps.

We recognize those Indian districts that show high COVID-19 cases have higher slum population concentration and lack of access to water supply within house premises. Our analysis (see Section 3, Fig. 2) shows several districts in Andhra Pradesh, Tamil Nadu, Chhattisgarh, Madhya Pradesh, Maharashtra, and Haryana are exposed to COVID-19 risk due to the large share of the population living in slums, where acute poverty, crowding, and unsanitary living environment are common characteristics. Another factor that potentially led to COVID-19 spread among districts in Rajasthan, Madhya Pradesh, Odisha, Meghalaya, Nagaland, Manipur, and Tripura is a high percentage of households without access to a water supply connection that limits their ability to hand wash frequently (Praharaj and Vaidya, 2020). We also identify districts that had a high burden of deaths from COVID-19, explained in the model by the percentage of elderly in a district and the share of women suffering from obesity and high blood sugar. The elderly population concentration is in the highest quantile (>5.2%) in 21 out of the 36 districts of Maharashtra. Kerala, Punjab, Karnataka, and Tamil Nadu also have more than half of their districts in this category. We find some similarities in the patterns of population health characteristics mapping between obesity and high blood sugar. In Kerala, Tamil Nadu, and Andhra Pradesh, over 28% of women face obesity, and more than 7.5% suffer from high blood sugar which has substantial implications for fatalities due to COVID-19 complications. Twelve of the 17 districts in Punjab, seven districts in Maharashtra, and all the areas of Delhi also display these severe obesity-related risks that might have shaped the pandemic outcomes in India. Our findings validate previous research findings (see Garg et al., 2020; Wu et al., 2020) that established a close association between pre-existing illness (comorbidities) and COVID-19 fatalities.

Our findings have long-term implications for policy, beyond the pandemic, as the significant factors of higher concentration of slum population and a higher share of households without access to the water supply have been, for decades, a major challenge for authorities in the rapidly urbanizing regions of the world. It is well established in the literature that limited services (e.g., water supply, solid waste management) are provided to the population living in slums, characterized by a lack of established land tenure and legal ownership of shelter. There is a perception among the authorities that providing public services to slums legitimizes the occupation (illegal) of public land. The effect of the COVID-19 pandemic on slum communities re-emphasizes the need for public agencies to address the vulnerability of the urban poor communities by re-imagining urban land use planning and infrastructure delivery models. Improved services for slums are critical to building long-term resilience among the urban poor communities to effectively fight disasters and pandemics. Considering the population living in slums faces extreme poverty, the issues of vaccine equity and affordability need emphasis. As of June 2021, the COVID-19 vaccines in India were free for people above 45 years of age. However, the cost of a single dose of vaccine in the private market in India is among the highest – between \$12 and \$17. There is an urgent need for governments to design targeted vaccination campaigns for at-risk communities and pockets of poverty to provide affordable or free vaccination to lift economically marginalized communities out of the pandemic.

The results from this study are also significant from medical geography and public health policy standpoint as a strong correlation between pre-existing health conditions (obesity, high blood sugar, and elderly population) with COVID-19 fatalities was established. The pressure on the healthcare system during the pandemic has spiked far beyond normal capacity, even in developed countries that spend nearly 20–40 times more on health care and have 3–6 times more hospital beds than India (Subramanian et al., 2020). The modeling of population health indicators and identification of high-risk districts reveal where future medical infrastructure capacities need to be developed. These findings have immediate implications for vaccination strategy for populations with comorbidities. It also urges a future discourse on the need for a new public health policy roadmap in India beyond the pandemic to address the issues of adequate and equitable distribution and access to public health infrastructure.

While our findings are novel, we recognize the need for interpretation with caution as the study is subject to some limitations. To the best of our knowledge, this investigation is the first to harness, and synthesize three different data sources, including the Census of India, COVID-19 India.org, and the DHS, to build a robust spatial regression model to characterize the nature of the pandemic in India. However, the temporal effect of these datasets is not always consistent; COVID-19 cases and deaths were reported as of May 9, 2021, the DHS population health data are from the 2015 to 2016 NFHS survey, and the Census demographic dataset was initially published in 2011. We acknowledge that demographic and household characteristics might have changed substantially since the last Census (2011); hence, the variables might have interacted differently with the pandemic outcomes if a more recent census dataset was available. The availability of timely and granular data on a standardized basis, especially in the LMICs like India, has been a challenge well documented in the literature. The heterogeneity in our model fit underscores the complexity of COVID-19 period prevalence, as the model, in some cases, could not fully explain the variances in the dataset (see Section 3). The reason for this could be that, in addition to the variables we examined, a number of other factors (e.g., social behavior, implementation of social distancing, clinical health infrastructure, climatic conditions, and government policies) all influence pandemic outcomes in large and diverse geographies such as India. The complexity and levels of decision-making influencing the spread and intensity of COVID-19 have not been fully unpacked.

There are also severe concerns about the efficiency and accuracy of government-provided data on COVID-19 cases and fatalities, as New York Times recently reported that in the most conservative scenario, the actual infection rate in India could be 26 times more than what is officially reported (Gamio and Glanz, 2021). There is also significant variation reported in the testing capacity across the states and a lack of accessibility to testing in a large rural population base in the country (Subramanian et al., 2020). The Indian government's ICMR-administered serosurvey estimated that nearly one in four individuals aged 10 years or older from the general population were exposed to COVID-19 by December 2020, amounting to 271 million infections in India (Murhekar et al., 2021b). In contrast, the official data confirmed only 10.4 million cases up to the end of 2020. Representative estimates of COVID-19 infection fatality rates by Cai et al. (2021) also indicate that India does not count deaths occurring when patients are at home. Such discrepancies mean that actual deaths are far exceeding official figures. In a country as large as India, even a small fluctuation in fatality rates could result in a difference of hundreds of thousands of deaths. Our study could achieve 70% accuracy in measuring and explaining the association between the chosen variables and COVID-19 cases and deaths even with these limitations and data gaps.

On the methodological side, we acknowledge that while GWR has high utility in epidemiology, particularly for infectious disease research and evaluations of health policies or programs, there are known limitations of the model, including problems of multicollinearity and the approaches to calculating goodness of fit statistics (Wheeler and Tiefelsdorf, 2005). Although GWR explicitly models spatial variation, the approach ignores potential dependencies among the local regression coefficients associated with different exogenous variables. This correlation of local regression coefficients can potentially invalidate any interpretation of individual GWR parameter estimates and can facilitate misleading conclusions (Kitchin and Thrift, 2009) if the situation is not properly diagnosed. Hence, we argue that GWR models should generally be utilized as tools for exploratory and descriptive data analysis and inferences should not be overplayed. Also, a new class of GWR models would be required to formally express the effects of these constraints, and emerging methods such as Conditional Autoregressive (CAR) models based on maximum likelihood and the Bayesian approach (De Oliveira, 2012) might provide a way out in this regard. A careful application of the technique and the use of diagnostic tools, including the ones applied in this study, such as local VIF inspection, scatter plots of the local parameter estimates, and maps of the local parameter correlation can also address the potential dependency issues in GWR research.

## Conclusion

The current article demonstrates the significance of using micro-level data and geospatial analysis to identify areas with potentially differential susceptibility to public health disasters, including the COVID-19. The study offers an approach to navigating the complexities and constraints associated with data availability in LMICs (see discussions in Section 4) to build robust spatial models for identifying the correlates of COVID-19 cases and deaths. It is important to recall that each of the six explanatory variables examined in this study demonstrated statistically significant locally varying associations with COVID-19 cases/deaths. We highlight how emerging sources of data, such as the DHS population health data in over 90 countries (IIPS and ICF International, 2017), can be leveraged to investigate novel questions that contribute to emergent research affecting the world's most vulnerable communities. As the governments and first-responders continue their efforts to contain the health crisis, our findings contribute to the epidemiologic understanding of geographic disparities and location-based risk factors for COVID-19, which can be used to inform a place-based and data-driven precision policy strategy. Outcomes from our analysis contribute to the growing body of knowledge on "health geography" (Dummer, 2008) by deconstructing health from a holistic perspective on society and space and conceptualizing the role of place, location and geography in health, well-being, and disease.

While previous spatial COVID-19 studies on India have concentrated on the state level (Gupta, Banerjee, and Das, 2020; Rafiq, Suhail, and Bazaz, 2020; Sarkar, Khajanchi, and Nieto, 2020), this is the first article to use spatial regression analysis to demonstrate the embeddedness and connectedness of districts, as well as the importance of relative locations to local decision-makers. We emphasize that what matters in the spread of COVID-19 is not only the contextual factors unique to a particular location, but also the unexposed features of its neighbors, and spatial regression-based approaches are critical for deciphering those dynamics. While the findings reported in this study are novel, they are not meant to be considered the only evidence, because various other factors, including income, clinical health infrastructure, air pollution, climate, and government policies, all influence the spatial distribution and trajectory of the pandemic. We share

this work as a proof-of-concept for further adaptation and testing with additional variables in order to start giving a clearer picture of the disease spread, which may be the scope of future research.

## Acknowledgements

The authors would like to recognize COVID-19India.org and the DHS program of the USAID for providing public data that are used in this research. The authors would also like to thank the reviewers for their constructive and detailed comments during various stages of peer review of this article.

## References

- Andersen, L. M., S. R. Harden, M. M. Sugg, J. D. Runkle, and T. E. Lundquist. (2021). "Analyzing the Spatial Determinants of Local Covid-19 Transmission in the United States." *Science of the Total Environment* 754, 142396. <https://doi.org/10.1016/j.scitotenv.2020.142396>
- Bhadra, A., A. Mukherjee, and K. Sarkar. (2021). "Impact of Population Density on Covid-19 Infected and Mortality Rate in India." *Modeling Earth Systems and Environment* 7, 623–9. <https://doi.org/10.1007/s40808-020-00984-7>
- Biswas, S. (2020). *India Coronavirus: "More Than Half of Mumbai Slum-dwellers Had Covid-19" – BBC News [WWW Document]*. BBC. Available at: <https://www.bbc.com/news/world-asia-india-53576653> [Accessed 18th November 2020].
- Cai, R., Novosad, P., Tandel, V., Asher, S., Malani, A. (2021). "Representative Estimates of COVID-19 Infection Fatality Rates from Three Locations in India." *medRxiv 2021.01.05.21249264*. <https://doi.org/10.1101/2021.01.05.21249264>
- Chi, G., and J. Zhu. (2019). *Spatial Regression Models for the Social Sciences*. California: SAGE Publications.
- Corsi, D. J., M. Neuman, J. E. Finlay, and S. v. Subramanian. (2012). "Demographic and Health Surveys: A Profile." *International Journal of Epidemiology* 41, 1602–13. <https://doi.org/10.1093/ije/dys184>
- COVID-19 India Org Data Operations Group. (2020). COVID-19 India Tracker [WWW Document]. Available at: <https://api.covid19india.org/> [Accessed 30th October 2020].
- De Oliveira, V. (2012). "Bayesian Analysis of Conditional Autoregressive Models." *Annals of the Institute of Statistical Mathematics* 64(1), 107–33.
- Desjardins, M. R., A. Hohl, and E. M. Delmelle. (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, 1–7. <https://doi.org/10.1016/j.apgeog.2020.102202>
- Dowd, J. B., L. Andriano, D. M. Brazel, V. Rotondi, P. Block, X. Ding, Y. Liu, and M. C. Mills. (2020). "Demographic Science Aids in Understanding the Spread and Fatality Rates of COVID-19." *Proceedings of the National Academy of Sciences of the United States of America* 117, 9696–8. <https://doi.org/10.1073/pnas.2004911117>
- Dummer, T. J. B. (2008). "Health Geography: Supporting Public Health Policy and Planning." *CMAJ* 178(9), 1177–80. <https://doi.org/10.1503/cmaj.071783>
- DuPre, N. C., S. Karimi, C. H. Zhang, L. Blair, A. Gupta, M. A. Lamyaa, M. A. Alharbi, W. Riten Mitra, P. McKinney, and B. Little. (2021). "County-Level Demographic, Social, Economic, and Lifestyle Correlates of COVID-19 Infection and Death Trajectories During the First Wave of the Pandemic in the United States." *The Science of the Total Environment* 786, 147495. <https://doi.org/10.1016/j.scitotenv.2021.147495>
- Fotheringham, A., and P. Rogerson. (2012). *The SAGE Handbook of Spatial Analysis*. California: SAGE Publications, Ltd. <https://doi.org/10.4135/9780857020130>
- Fotheringham, A. S. (2012). "Geographically Weighted Regression." In *The SAGE Handbook of Spatial Analysis*, 242–53. California: SAGE Publications, Ltd. <https://doi.org/10.4135/9780857020130.n13>
- Franch-Pardo, I., M. Desjardins, I. Barea-Navarro, and A. Cerdà. (2021). "A Review of GIS Methodologies to Analyze the Dynamics of COVID-19 in the Second Half of 2020." *Transactions in GIS* 25(5), 2191–239. <https://doi.org/10.1111/tgis.12792>

- Franch-Pardo, I., B. M. Napoletano, F. Rosete-Verges, and L. Billa. (2020). "Spatial Analysis and GIS in the Study of COVID-19. A Review." *Science of The Total Environment* 739, 140033. <https://doi.org/10.1016/j.scitotenv.2020.140033>
- Gamio, L. and Glanz, J. (2021). Just How Big Could India's True Covid-19 Case and Death Toll Be? [WWW Document]. The New York Times. Available at: <https://www.nytimes.com/interactive/2021/05/25/world/asia/india-covid-death-estimates.html> [Accessed 29th May 2021].
- Garg, S., L. Kim, M. Whitaker, A. O'Halloran, C. Cummings, R. Holstein, et al. (2020). "Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 – COVID-NET, 14 States, March 1–30, 2020." *MMWR. Morbidity and Mortality Weekly Report* 69, 458–64. <https://doi.org/10.15585/mmwr.mm6915e3>
- Glass, G. E. (2000). "Update: Spatial Aspects of Epidemiology: The Interface with Medical Geography." *Epidemiologic Reviews* 22, 136–9.
- Gupta, A., S. Banerjee, and S. Das. (2020). "Significance of Geographical Factors to the COVID-19 Outbreak in India." *Modeling Earth Systems and Environment* 6, 2645–53. <https://doi.org/10.1007/s40808-020-00838-2>
- Hamidi, S., R. Ewing, and S. Sabouri. (2020). "Longitudinal Analyses of the Relationship between Development Density and the COVID-19 Morbidity and Mortality Rates: Early Evidence From 1,165 Metropolitan Counties in the United States." *Health and Place* 64, 102378. <https://doi.org/10.1016/j.healthplace.2020.102378>
- IIPS & ICF International (2017). *National Family Health Survey (NFHS-4), 2015–16: India*. Mumbai: IIPS.
- Joe, W., A. Kumar, S. Rajpal, U. S. Mishra, and S. V. Subramanian. (2020). "Equal Risk, Unequal Burden? Gender Differentials in COVID-19 Mortality in India." *Journal of Global Health Science* 2, 1-15. <https://doi.org/10.35500/jghs.2020.2.e17>
- Kapitsinis, N. (2020). "The Underlying Factors of the COVID-19 Spatially Uneven Spread. Initial Evidence from Regions in Nine EU Countries." *Regional Science Policy and Practice* 12, 1027–45. <https://doi.org/10.1111/rsp3.12340>
- Kitchin, R., and N. Thrift. (2009). *International Encyclopedia of Human Geography*. Amsterdam: Elsevier.
- Mollalo, A., B. Vahedi, and K. M. Rivera. (2020). "GIS-Based Spatial Modeling of COVID-19 Incidence Rate in the Continental United States." *Science of the Total Environment* 728, 138884. <https://doi.org/10.1016/j.scitotenv.2020.138884>
- Murhekar, M. V., T. Bhatnagar, S. Selvaraju, V. Saravanakumar, J. W. V. Thangaraj, N. Shah, et al. (2021a). "SARS-CoV-2 Antibody Seroprevalence in India, August–September, 2020: Findings from the Second Nationwide Household Serosurvey." *The Lancet Global Health* 9, e257–66. [https://doi.org/10.1016/S2214-109X\(20\)30544-1](https://doi.org/10.1016/S2214-109X(20)30544-1)
- Murhekar, M. V., T. Bhatnagar, J. W. V. Thangaraj, V. Saravanakumar, M. S. Kumar, S. Selvaraju, et al. (2021b). "SARS-CoV-2 Sero-Prevalence Among General Population and Healthcare Workers in India, December 2020–January 2021." *SSRN Electronic Journal*. 1-29. <https://doi.org/10.2139/ssrn.3797589>
- Nakaya, T., Charlton, M., Lewis, P., Brunson, C., Yao, J., and Fotheringham, S. (2012). GWR4 User Manual. Windows Application for Geographically Weighted Regression Modelling.
- ORGI. (2020). *District Census Hand Book*. Ministry of Home Affairs, Government of India. New Delhi: Ministry of Home Affairs, Government of India.
- Perone, G. (2021). "The Determinants of COVID-19 Case Fatality Rate (CFR) in the Italian Regions and Provinces: An Analysis of Environmental, Demographic, and Healthcare Factors." *Science of the Total Environment* 755, 142523. <https://doi.org/10.1016/j.scitotenv.2020.142523>
- Pfeiffer, D. U., T. P. Robinson, M. Stevenson, K. B. Stevens, D. J. Rogers, and A. C. A. Clements. (2008). *Spatial Analysis in Epidemiology*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198509882.001.0001>
- Popovich, N., Singhvi, A., and Conlen, M. (2020). Where Chronic Health Conditions and Coronavirus Could Collide. The New York Times.
- Praharaj, S., and H. Han. (2019). "Building a Typology of the 100 Smart Cities in India." *Smart and Sustainable Built Environment* 8(5), 400-14. <https://doi.org/10.1108/SASBE-04-2019-0056>

- Praharaj, S., J. H. Han, and S. Hawken. (2018). "Evolving a Locally Appropriate Indicator System for Benchmarking Sustainable Smart Cities in India." In *World Sustainability Series*, 253–74. Switzerland: Springer. [https://doi.org/10.1007/978-3-319-73293-0\\_15](https://doi.org/10.1007/978-3-319-73293-0_15)
- Praharaj, S. and Vaidya, H. (2020). *The Urban Dimension of COVID-19 in India: COVID Outbreak and Lessons for Future Cities*. New Delhi.
- Praharaj, S., and H. Han. (2021). "Human mobility impacts on the surging incidence of COVID-19 in India." *Geographical Research* 60(1), 18–28. <https://doi.org/10.1111/1745-5871.12502>
- Praharaj, S., D. King, C. Pettit, and E. Wentz. (2020). "Using aggregated mobility data to measure the effect of COVID-19 policies on mobility changes in Sydney, London, Phoenix, and Pune." *Findings*. <https://doi.org/10.32866/001c.17590>
- Rafiq, D., S. A. Suhail, and M. A. Bazaz. (2020). "Evaluation and Prediction of COVID-19 in India: A Case Study of Worst Hit States." *Chaos, Solitons and Fractals* 139, 110014. <https://doi.org/10.1016/j.chaos.2020.110014>
- Sannigrabi, S., Pilla, F., Basu, B., and Basu, A.S. (2020). "The Overall Mortality Caused by COVID-19 in the European Region is Highly Associated with Demographic Composition: A Spatial Regression-based Approach." *arXiv*. <https://doi.org/10.1016/j.scs.2020.102418>
- Sarkar, K., S. Khajanchi, and J. J. Nieto. (2020). "Modeling and Forecasting the COVID-19 Pandemic in India." *Chaos, Solitons and Fractals* 139, 1-16. <https://doi.org/10.1016/j.chaos.2020.110049>
- Subramanian, S. V., O. Karlsson, W. Zhang, and R. Kim. (2020). "Geo-mapping of COVID-19 Risk Correlates Across Districts and Parliamentary Constituencies in India." *Harvard Data Science Review* 1, 1-39. <https://doi.org/10.1162/99608f92.68bb12e4>
- Sun, F., S. A. Matthews, T. C. Yang, and M. H. Hu. (2020). "A Spatial Analysis of the COVID-19 Period Prevalence in U.S. Counties through June 28, 2020: Where Geography Matters?" *Annals of Epidemiology* 52, 54–59.e1. <https://doi.org/10.1016/j.annepidem.2020.07.014>
- Wang, C., Z. Li, M. Clay Mathews, S. Praharaj, B. Karna, and P. Solís. (2020). "The Spatial Association of Social Vulnerability with COVID-19 Prevalence in the Contiguous United States." *International Journal of Environmental Health Research*. 32(5), 1147-54. <https://doi.org/10.1080/09603123.2020.1847258>
- Wheeler, D., and M. Tiefelsdorf. (2005). "Multicollinearity and Correlation among Local Regression Coefficients in Geographically Weighted Regression." *Journal of Geographical Systems* 7(2), 161–87.
- WHO (2021). WHO Coronavirus Disease (COVID-19) Dashboard [WWW Document]. Available at: <https://covid19.who.int/> [Accessed 26th July 2021].
- World Bank (2020). Projected Poverty Impacts of COVID-19 (coronavirus).
- Worldometer (2021). COVID-19 Coronavirus Pandemic [WWW Document]. Available at: <https://www.worldometers.info/coronavirus/> [Accessed 26th July 2021].
- Wu, C., X. Chen, Y. Cai, J. Xia, X. Zhou, S. Xu, et al. (2020). "Risk Factors Associated with Acute Respiratory Distress Syndrome and Death in Patients with Coronavirus Disease 2019 Pneumonia in Wuhan, China." *JAMA Internal Medicine* 180, 934–43. <https://doi.org/10.1001/jamainternmed.2020.0994>
- Zhang, J., H. Li, M. Lei, and L. Zhang. (2021). "The Impact of the COVID-19 Outbreak on the Air Quality in China: Evidence from a Quasi-Natural Experiment." *Journal of Cleaner Production* 296, 126475. <https://doi.org/10.1016/j.jclepro.2021.126475>
- Zhang, W., R. Kim, and S. v. Subramanian. (2020). "Crowdsourcing for Robust, Real-Time COVID-19 Data." *Nature India*. <https://doi.org/10.1038/nindia.2020.109>
- Zhou, F., T. Yu, R. Du, G. Fan, Y. Liu, J. Xiang, et al. (2020). "Clinical Course and Risk Factors for Mortality of Adult Inpatients with COVID-19 in Wuhan, China: A Retrospective Cohort Study." *The Lancet* 395, 1054–62. [https://doi.org/10.1016/S0140-6736\(20\)30566-3](https://doi.org/10.1016/S0140-6736(20)30566-3)

## Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web site.