IJSA (2024), Vol. 8, No. 1, 14-36
DOI: https://doi.org/10.29244/ijsa.v8i1p14-36

Indonesian Journal of Statistics and Its Applications
Available at https://journal-stats.ipb.ac.id/index.php/ijsa
e-ISSN: 2599-0802

Addressing Multicollinearity in Spatial Modelling: A District Level Spatial Analysis of Pandemic COVID-19 in India*

Shalini Chandra [1], Megha Sharma 2[2‡]

[1,2] Department of Mathematics and Statistics, Banasthali Vidyapith,
Rajasthan, India-304022.
‡corresponding author: meghasharma15aug@gmail.com

Abstract

This study focuses on conducting spatial analysis of COVID-19 at the district level in India. Leveraging data from www.covidindia.org for confirmed cases and deaths, and integrating population characteristics from the National Family Health Survey 5 (2019-2021) and supplementary sources. The objective is to identify risk factors using spatial modelling techniques while addressing multicollinearity through principal component analysis (PCA). This study utilizes spatial analysis to identify COVID-19 hotspots and cold spots at the district level in India. It highlights highly affected districts such as Mumbai, Pune, Chennai, Kolkata, and Bengaluru, as well as low affected districts in central and north-eastern regions. The study utilized the spatial lag model (SLM), spatial error model (SEM), geographical weighted regression (GWR), and multiscale geographical weighted regression (MGWR) models to analyze the impact of demographic, socioeconomic, climatic, and comorbidity factors on COVID-19, accounting for spatial proximity. Among these models, MGWR exhibited superior performance. Key risk factors associated with the COVID-19 phenomenon identified, providing insights into the impact of household conditions, educational level of women, tobacco and alcohol consumption rates, number of health centres, and climatic factors. Moreover, the local coefficients estimated by MGWR model furnish detailed information regarding the strength and direction of the relationships between predictors and COVID-19 cases and deaths within each spatial unit. The findings emphasize the significance of addressing multicollinearity in spatial modelling. It is beneficial for accurate parameter estimation, proper interpretation of coefficients, improved spatial analysis, and providing reliable insights to support decision-making in spatial contexts.

Keywords: covid-19, multicollinearity, principal component analysis, spatial analysis.

* Received: Sep 2023; Reviewed: Oct 2023; Published: Jun 2024

1. Introduction

Many previous research found that the existence of social inequalities can contribute to circumstances that make it easier for the disease to spread, thereby making it more difficult to control the pandemic (Ahmed et al., 2020). Poor living conditions (Pereira & Oliveira, 2020), population density (Gupta et al., 2020), inadequate access to healthcare, and a large population of susceptible population, such as the older and those with existing medical conditions (Dutta et al., 2021), are all factors that make any region vulnerable to the spread of the virus. In addition, the severity of COVID-19 in China was found to have a positive association with temperature (Chen et al., 2020). A comparable relationship between temperature and COVID19 cases has been discovered in various countries, including India (Gupta et al., 2020), Indonesia (Tosepu et al., 2020), and on a global scale (Chen et al., 2020). Some other factors like the prevalence of slums within cities (Sridhar, 2023), smoking habits (Puebla Neira et al., 2021), and many more contribute to an increased risk of transmission and disparities in access to prevention and treatment measures.

Infectious disease transmission is linked to geographical proximity, as the transmission is more likely if at-risk individuals are close in spatial proximity. Incorporating the spatial dimension into epidemiological investigations allows for more informative descriptive analysis and the acquisition of additional insights into the causal process under study (Mollalo & Khodabandehloo, 2016; Pfeiffer et al., 2008). Spatial models have become vital in examining and interpreting the spread of COVID-19 channelling. The spatial distribution of COVID-19 in Iran, Bangladesh, and some European countries has been studied (Adekunle et al., 2020; Dutta et al., 2021; Sarkar et al., 2021). They used various spatial models, such as spatial lag model (SLM), spatial error models (SEM), geographically weighted regression (GWR), and multi-scale geographical weighted regression (MGWR) to determine the clustered regions with high or low COVID-19 incidence in different countries.

Spatial models have demonstrated their utility in comprehending and assessing pandemic dynamics. However, there are instances where the presence of correlated risk factors can lead to the challenge of multicollinearity when using these models. Multicollinearity refers to the presence of high correlation among independent variables in a regression analysis, restrict the precision of parameter estimates and make it challenging to discern the individual effects of variables. Several research papers have highlighted the issue of multicollinearity while employed spatial models for pandemic analysis. For example, a study which is focused on spatial modelling of dengue transmission in Mexico. The authors found that including various spatial variables, such as population density and climatic factors, resulted in high multicollinearity among these variables. This multicollinearity limited the ability to accurately estimate the effects of individual variables and affected the model's predictive power (Dzul-Manzanilla et al., 2021). Another study examined the spatial distribution of COVID-19 in China. The authors utilized a spatial autoregressive model to capture the spatial dependence among neighbouring regions. However, they encountered multicollinearity issues when including multiple socioeconomic variables, such as population density, income level, and healthcare resources. The high correlation among these variables made it challenging to ascertain their distinct influences on COVID-19 transmission (Wang et al., 2021). Various approaches have

been suggested by researchers to address this issue, with Principal Component Analysis (PCA) emerging as a popular choice. PCA is employed to convert interrelated variables into uncorrelated components, thereby diminishing multicollinearity while retaining essential information (Shen & Zhu, 2015).

This study conducted a specific analysis for each wave of the COVID-19 pandemic outbreak in India to examine the impact of district-level vulnerabilities on the spread of the virus. Spatial proximity was taken into account in assessing the influence of independent factors on the spatial dispersion of COVID-19 during the first two waves. Four spatial models- SLM, SEM, GWR, MGWR, were utilized to capture spatial dependencies and variations in the relationships between independent variables and COVID-19 outcomes. Although Spatial models offer valuable insights into the behavior of pandemics; they frequently encounter the issue of multicollinearity. To address this challenge, PCA can be employed as an appropriate technique to mitigate multicollinearity in spatial models. Through the utilization of PCA in the spatial models, this research has the potential to successfully mitigate the issue of collinearity among variables, resulting in improved accuracy and interpretability of their models. The findings from this analysis can support public health officials and policymakers in identifying hotspots, monitoring the spread, identifying disparities, and evaluating the effectiveness of interventions.

2. Methodology

This study used statistical and geographical techniques to analyze the geographic distribution of diseases and health outcomes and to identify patterns, relationships, and risk factors that can inform public health planning and decision-making. A common framework for conducting spatial analysis in epidemiology involves integrating geographic information with epidemiological data to examine the spatial distribution of disease and the relationship between disease and geographic variables. Figure 1 is a diagrammatic representation of conceptual framework adapted for this study.

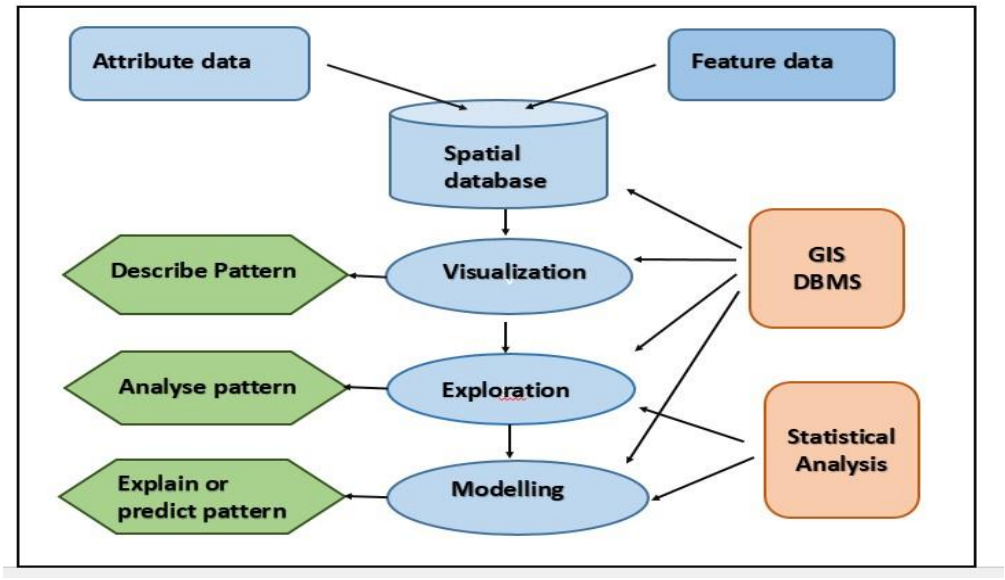


Figure 1: Framework of this study

The specific analytic objectives lead to three analytical methods: visualization, exploration, and modelling. The first two groups cover techniques solely concerned with examining data's spatial dimension and modelling explored or predicting cause and effect associations using both spatial and non-spatial data sources. Our study included districts from all states and union territories in India, except six states (Assam, Delhi, Goa, Manipur, Telangana, Sikkim) with no COVID-19 updates at the district level in the state bulletin. This study took into account district boundaries as of 2019 (figure 2 provides the district's name and boundaries, which aid in understanding all the maps presented in the study).

We extracted district-level data on daily confirmed cases of COVID-19 and associated deaths in India for this study from the website www.covidindia.org. This public domain collects data through state bulletins and official handles. They halted the operation after 18 months of daily updates. As a result, this study limits the analysis of data till October 2021.

Figure 2: India administrative boundaries as of 2019

Source: googlesand.blogspot.com

2.1 Data Collection and Preparations

So far, several variables have impacted COVID-19 spread during these pandemic outbreaks, from which some of the essential independent variables that may have affected COVID-19 spread in Indian districts have been selected. Table 1 lists these independent variables, their descriptions containing the reason behind taking these variables in our study, and the sources from which they were derived. India has experienced multiple waves of COVID-19 since the pandemic began. It is worth noting

that these waves need to be clearly defined. India has undergone several surges of COVID-19 since the onset of the pandemic. Specifically, India encountered two distinct waves of COVID-19 between December 2019 and October 2021. These waves occurred during the periods of March 2020 to December 2020 and January 2021 to October 2021, respectively.

Table 1: List of independent variables, descriptions/ justifications and the sources.

S.no.	Indicators/ Abbreviation	Assumptions/ Justifications	Data sources
1	Population living in households that use an improved sanitation facility (ISF)	The environment in which people live plays a significant role in the transmission of COVID-19. Factors such as overcrowding, sanitation and hand hygiene all contribute to susceptibility and should not be overlooked.	National Family Health Survey (NFHS-5) (2019-21) (District factsheet)
2	Households using clean fuel for cooking (CF)		
3	Population living in households with electricity (Elec)		
4	Population below age 15 years (age-15)	Older population have higher risk of death after infected.	
5	Households with any usual member covered under a health insurance/ financing scheme (HI)	Accessible healthcare systems, affordability, capacity, and health security are vital for managing epidemics and promoting treatment-seeking.	
6	Women who are literate (LW)	Women's literacy empowers them with knowledge, enabling them to understand COVID-19 prevention, access reliable information, and make informed decisions.	
7	Educated women with 10+ years of schooling. (LW 10+)		
8	Adults' blood sugar levels (age 15+).	Blood Sugar Level and Hypertension among Adults (age 15+) may regulate the severity of COVID-19 cases.	
9	Hypertension among Adults (age 15+) (HT)		
10	Tobacco uses among those 15+ (TP)	Smoking or tobacco or any kind of alcohol being exposed in any form can reduce the risk of COVID-19 infection (WHO 2020).	
11	Alcohol use among those 15+. (AP)		
12	Population Density (density)	High population density and urban areas posing a higher risk for the spread of the highly contagious SARS-CoV-2 virus.	Office of the Registrar General of India
13	Proportion of urban population (urban)		

14	Health Center [Sub center + PHCs+ CHCs] (HC)	Higher population per healthcare institution indicates lower resilience in dealing with COVID-19.	Rural Health statistics
15	Average temperature (temp)	The severity of COVID-19 is found to have a positive relationship with temperature	NASA open data portal
16	Proportion of poor population(poor)	Studies have shown that areas with high poverty rates tend to have higher rates of COVID-19 infections.	Global Data Lab

2.2 Study Area and Descriptive Statistics

The study area for this research is India, located in South Asia and positioned geographically between latitudes 8°4' and 37°6' north and longitudes 68°7' and 97°25' east. India shares its borders with Pakistan to the northwest, China and Nepal to the north, Bhutan to the northeast, and Bangladesh and Myanmar to the east, while the Indian Ocean bounds it to the south. India boasts diverse landscapes, ranging from the Himalayan Mountain range in the north to coastal regions in the south, experiencing various climates, including tropical, subtropical, arid, and alpine, contributing to agricultural and ecological diversity. As the world's second-most populous country with over 1.3 billion people, India is administratively divided into districts, totaling 640 as of 2019. In the context of the first two waves of COVID-19, this research focuses on 16 dependent variables and two independent variables—TCC and deaths—utilizing data from 626 districts (due to data unavailability), resulting in a comprehensive dataset of approximately 10,000 observations for spatial and statistical analysis.

In the initial and subsequent phases of the pandemic, specific Indian districts, including Bangalore, Mysuru, Belagavi, Pune, Mumbai, Thane, Nagpur, Ernakulam, Malappuram, Nashik, Kollam, Kolkata, Chennai, Coimbatore, Chittoor, among others in Kerala, Tamil Nadu, Andhra Pradesh, and West Bengal, experienced heightened COVID-19 cases and deaths. Geographical variations were evident, with northern and central states like Lucknow, Varanasi, Kanpur, Jaipur, Jodhpur, Ludhiana, and Jalandhar significantly affected, while areas like Hathras, Mahoba, Burhanpur, Agar Malwa, Mandla, and Baranala reported fewer cases. Generally, central and northeastern regions had lower confirmed cases and deaths in both waves.

The data indicates higher population density in Bihar, West Bengal, and Kerala, with 29 districts among the top 10%. On average, 4.24% of the population in these districts is aged 65 and above. Notably, Maharashtra, Kerala, Karnataka, Goa, and Punjab show a significant prevalence of districts with an aging population. Specifically, 15 out of Maharashtra's 36 districts and 9 out of Kerala's 14 districts rank in the top 10% for the percentage of elderly population. On average, 20.19% of households in Indian districts lack water supply within their premises. The data reveals pronounced water supply challenges in numerous districts of Odisha, Madhya Pradesh, and Rajasthan. Additionally, 10 districts in Andhra Pradesh and 5 in Maharashtra fall into this category. Kerala, Goa, Tamil Nadu, and Andhra Pradesh also exhibit a significant presence of districts with the highest percentages (>7.5%) of the population facing

elevated blood sugar levels.

The data highlights specific districts in Rajasthan, such as Jaisalmer and Barmer, known for extremely high temperatures. Districts in the northern plains, including parts of Uttar Pradesh, Bihar, and Haryana, may also experience high temperatures. Gujarat, Maharashtra, and certain parts of Kerala might encounter high humidity levels. Notably, alcohol and tobacco consumption are high in districts of northeastern states, Punjab, Goa, and select districts in Rajasthan.

According to the National Family Health Survey (NFHS-5), approximately 41% of India's total population has at least one member enrolled in health insurance or a health scheme. Rajasthan and Andhra Pradesh lead with the highest proportions of households covered (88% and 80%, respectively), while the Andaman and Nicobar Islands and Jammu and Kashmir show the lowest coverage, each below 15%.

2.3 Visualization

By using visualization techniques, patterns and discrepancies in the data can be identified. The most widely used approach for visualizing this type of data is through choropleth maps that employ quantile breaks. These maps use various colors to depict the intensity of variables of interest in each geographic region. With quantile breaks, the data distribution is divided into five segments: extremely low, low, medium, high, and extremely high. Visualization allows for the identification of patterns and errors in the data.

2.4 Exploration

Spatial data exploration involves the application of statistical methods to determine whether observed spatial patterns are random. Cluster analysis can be performed using either nonspecific (global) or specific (local) techniques. In order to use statistical methods that take spatial dependencies into account, a spatial weight matrix must be generated to describe how observations in a dataset are related to each other. In this study, a queen contiguity matrix was used, where neighbouring units are defined as sharing a common boundary or vertex. Global techniques are utilized to determine the presence of clustering across the entire study area but do not provide precise information on cluster locations. These methods generate a single statistic that quantifies the degree of spatial clustering, which can be evaluated for statistical significance. The most commonly used measure for this purpose is Moran’s I statistic, which is a widely accepted indicator of global spatial autocorrelation. Moran’s, I calculate global spatial autocorrelation among observations and ranges from -1 to 1. Negative Moran’s I value indicate dispersion (clustering of dissimilar values), positive values indicate clustering (clustering of similar values), and values close to zero represent absolute spatial randomness, i.e., no autocorrelation (Tu & Xia, 2008). However, Moran’s I statistic is incapable of detecting structural instability in the dataset. To identify spatial non-stationarity or locations of outliers, the LISA (local indicators of spatial association) tool was utilized to calculate local spatial autocorrelation. This tool describes significant correlations at specific locations as local spatial clusters (hot spots) or correlations between observations and neighbouring observations (Anselin, 1995).

2.5 Modelling

This section incorporates regression modelling to measure the impact of independent variables on the spatial dispersion of a specific outcome. The assumption of independence of observations is fundamental to many classical statistical methods. If a dataset exhibits spatial dependencies, this assumption is violated. Incorporating the spatial dimension into epidemiological investigations makes conducting more informative descriptive analyses and obtaining further insights into the underlying causal processes possible. In this study, spatial regression models were employed to measure the impact of independent variables on the spatial dispersion of a particular outcome, taking into account spatial dependencies. The spatial regression models included: SLM, SEM, GWR, and MGWR.

Multicollinearity poses a common challenge in utilizing spatial models to gain insights into pandemic behavior. To tackle this issue, the present study employed PCA as a means of addressing multicollinearity. PCA is a statistical technique commonly employed to uncover underlying latent factors within a dataset. PCA aims to reduce the dimensionality of the original variables by identifying linear combinations, known as principal components, that capture the maximum amount of variation in the data. By extracting these components, which are ordered based on the amount of variance they explain, PCA allows for a more concise representation of the data while retaining as much information as possible. This technique is particularly useful in mitigating multicollinearity issues in spatial models, as it helps identify independent factors that contribute significantly to the overall variance, thus aiding in the interpretation and understanding of the relationships between variables. To facilitate additional analysis, these components obtained were employed as independent variables in the spatial
The SLM (eq-1) was used to estimate the impact of independent variables on the dependent variable while accounting for spatial dependency. The SEM (eq-2) extended the classical regression model and accounted for spatial dependence in the disturbance term. The SLM model with n number of observations, and m number of independent variables presented in equation 1.

$$y = \rho W y + X\beta + \epsilon \qquad (1)$$

where y as the $n \times 1$ vector of dependent variable, X as the $n \times m$ matrix of independent variables, β as the vector of regression coefficients, the spatial autocorrelation coefficient of y represented by ρ, W as spatial weight matrix and ϵ is random error. Equation 2 presented SEM model with μ vector of spatially dependent disturbance terms, and λ its spatial autocorrelation coefficient.

$$y = X\beta + \mu, \mu = \lambda W \mu + \epsilon \qquad (2)$$

The SLM and SEM models assumed spatial stationarity, meaning that the relationships between dependent and independent variables did not vary across space. In contrast, the GWR model (eq-3) estimated local interactions between the dependent and independent variables by fitting a regression model to each feature in the dataset (Oshan et al., 2019). Finally, the MGWR model was an extension of the

GWR model that studied the relationships of independent and dependent variables at different spatial scales by using varying bandwidths to define the neighbourhood around each feature rather than a single, constant bandwidth for the entire study area (Fotheringham et al., 2017). The GWR model is presented by equation 3 given below

$$y = \sum_{j=1}^{m} X_{ij}\beta_{ij} + \in_i \qquad i = 1,2,3, \ldots, n. \qquad (3)$$

and parameters estimates for each independent variable at i^{th} location is given by

$$\hat{\beta}(i) = (X'W(i)X)^{-1}\,(X'W(i)y) \qquad (4)$$

where $\hat{\beta}(i)$ is $m \times 1$ vector of parameter estimates, $W(i)$ is spatial weight matrix calculated by the Gaussian kernel function and the bandwidth which is based on Euclidean distance. MGWR model presented in equation 5 with β_{bwj} as the bandwidth used for calibration of the j^{th} relationship.

$$y_i = \sum_{j=1}^{m} X_{ij}\beta_{bwj} + \in_i \quad i = 1,2,3, \ldots, n. \qquad (5)$$

The analysis on SEM, SLM, spatial association (Global Moran's I and LISA) was done using GeoDa software while the GWR and MGWR model was done using MGWR 2.2.1 software. R^2 and AIC were used to compare the performances of various models, explaining total COVID-19 confirmed cases and deaths. R-square measures the goodness of fit; its values range from 0 to 1. Furthermore, Akaike information criteria (AIC) is a model performance measure that can compare predictive models while accounting for model complexity. The model with lower AIC value and higher value of R^2 better fits the observed data.

3. Results

3.1 Visualization & Exploration

This study employed choropleth maps using quantile breaks to visualize the total confirmed cases and total deaths during the pandemic outbreak, yielding successful results. Referring to Figure 3, the districts that exhibited the highest numbers of confirmed COVID19 cases and deaths were Bangaluru, Mysuru, Belagavi, and 13 other districts in Karnataka.

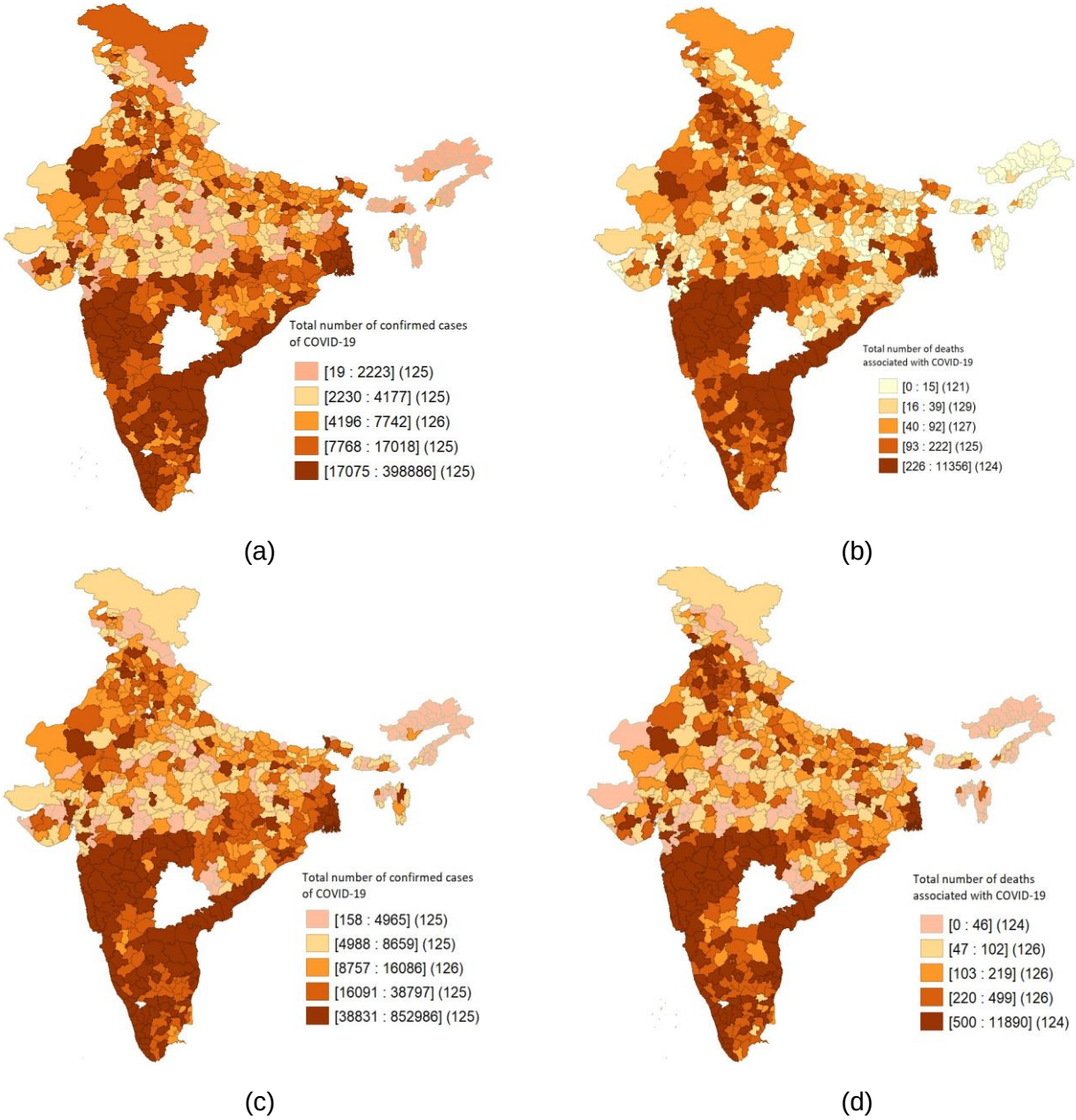


Figure 3: Quantitative spatial distribution of (a, c) total number of confirmed cases of COVID19, (b, d) total number of deaths associated with COVID-19 in 1st wave and 2nd wave respectively in Indian districts.

Additionally, in Maharashtra, the districts of Pune, Mumbai, Thane, Nagpur, and 29 out of 35 districts stood out. Similar trends were observed in Kerala, Tamil Nadu, Andhra Pradesh, and West Bengal, particularly in districts such as Ernakulam, Malappuram, Nashik, Kollam, Kolkata, Chennai, Coimbatore, Chittoor, and their adjacent districts. These districts were among the most affected during the entire duration of the pandemic analyzed in this study.

There were marked geographical distinctions among the northern and central states of India, with some districts like Lucknow, Varanasi, Kanpur, Jaipur, Jodhpur, Ludhiana and Jalandhar experiencing a high level of contagion while other areas like Hathras, Mahoba, Burahnpur, Agar Malwa, Mandla and Baranala and the locations

around them having a much lesser effect. In contrast, the central and northeastern regions districts of had the fewest confirmed cases and deaths in both waves.

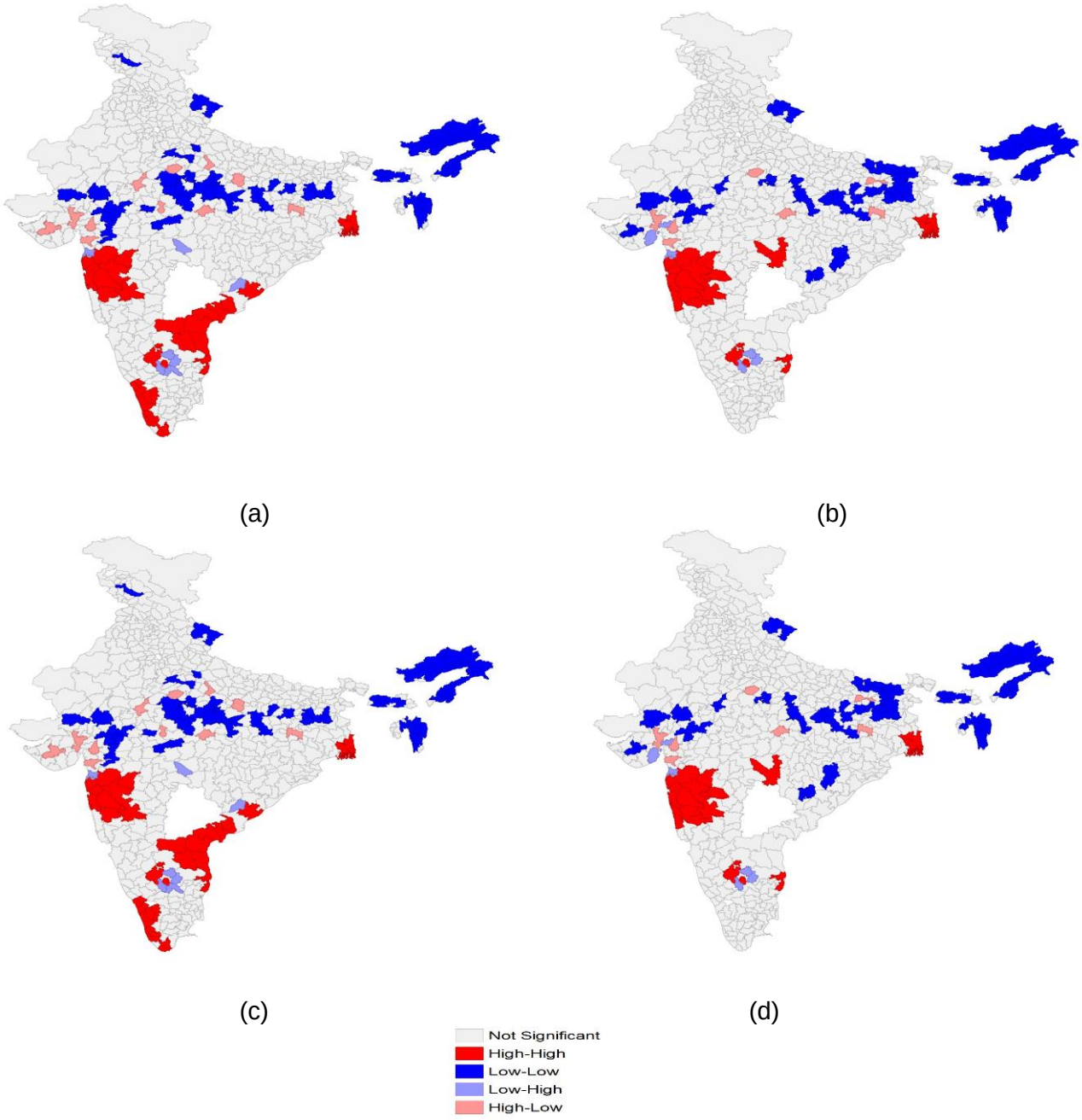


Figure 4: LISA clusters of (a, c) total number of confirmed cases of COVID-19, (b, d) total number of deaths associated with COVID-19 in 1st wave and 2nd wave respectively in Indian districts.

The global Moran's I statistic values for cumulative confirmed cases and deaths due to COVID-19 were significant for both waves (0.31, 0.43, and 0.27, 0.43, respectively, with p-value< 0.05), indicating strong spatial autocorrelation among Indian districts. Further, the LISA tool was employed to identify significant local clustering and detect non-clustered areas within the study that may be missed by global tests.

Using the LISA tool, the study found that the districts with the highest concentration of confirmed cases and deaths during both waves were the same, including Maharashtra, Kerala, Andhra Pradesh, West Bengal, and Karnataka. In contrast, the northern and central regions exhibited low clustering during the first wave, and the central region was also identified as having low clustering in the second wave (see figure-4) and only a few districts fell into the high-low and low-high clusters.

3.2 Modelling

3.2.1. Multicollinearity

Condition index, Variance Inflation Factor (VIF) and tolerance offer valuable information regarding the existence and intensity of multicollinearity in a regression model. High condition index, high VIF, and low tolerance suggest a high degree of multicollinearity among independent variables.

Table 2: Summary table provides the VIF and tolerance values for all the independent variables

Variables	Tolerance	VIF
Age (-15)	0.361	2.772
Elec	0.637	1.57
ISF	0.382	2.619
CF	0.298	3.359
HI	0.666	1.502
WL	0.172	5.799
WL (10+)	0.187	5.34
TP	0.315	3.173
AP	0.619	1.616
density	0.765	1.308
urban	0.396	2.527
HC	0.954	1.048
temp	0.572	1.748
poor	0.347	2.88
HT	0.55	1.817
diab	0.423	2.365

Table 3: Summary table provides the condition Index (CI) values for all the dimensions of the model.

Model dimension	1	2	3	4	5	6	7	8	9
CI	1	4.77	5.04	5.29	5.39	7.4	12.5	12.74	12.89
Model dimension	10	11	12	13	14	15	16	17	
CI	13.27	13.36	13.7	24.21	24.79	25.31	25.44	27.96	

In the present study, the table 2&3 reveals that all independent variables have VIF values ranges 1 to 6, tolerance values less than 0.8 and condition index nearly 30. This pattern clearly indicates a moderate level of multicollinearity among the variables,

emphasizing the challenges in discerning their distinct effects. In order to mitigate multicollinearity, this study employed PCA, by extracting a smaller set of uncorrelated variables, referred as components. The suitability of the data for PCA was assessed using measures such as the Kaiser-Meyer-Olkin (KMO) test or Bartlett's test of sphericity. These tests (KMO=0.779, and p-value for Bartlett's test of sphericity=0.0001) provided insights into the appropriateness of applying factor analysis to the dataset, ensuring that the assumptions and requirements of the analysis were met.

Based on the Kaiser criterion, which suggests retaining principal components with eigenvalues exceeding 1 as they account for substantial variation in the original dataset, five components were retained for further analysis. To aid in the interpretation of the PCA, a varimax rotation was applied. The rotated component matrix was examined, and variables with a loading threshold of 0.7 were considered to have the most influence on each component.

Table 4: Loadings of the varimax rotated components

Rotated Component Matrix					
Variables	Component				
	1	2	3	4	5
Age (-15)	**-.801**	-.171	-.077	.136	.075
Elec	.551	.134	-.295	-.349	-.178
ISF	**.765**	-.117	.119	-.046	.055
CF	**.773**	.206	-.208	.230	-.001
HI	.064	.308	.239	**-.744**	-.034
LW	**.846**	-.009	.096	.041	-.003
LW (10+)	**.855**	.043	.102	.203	.098
TP	-.670	-.299	**.719**	-.038	-.149
AP	-.020	-.110	**.898**	-.143	-.059
density	.105	.429	.048	.617	-.118
urban	.580	.449	-.143	.280	-.095
HC	.075	.015	-.009	-.034	**.949**
temp	-.220	**.801**	-.161	-.129	-.002
poor	**-.796**	.156	-.048	.212	-.046
HT	.392	.643	.108	.026	.069
diab	.576	.171	.499	.025	.176

Referring to the information presented in Table 4, it can be observed that the first component exhibited high loadings on variables associated with age, poor housing conditions, and education level among women (Age (-15), ISF, CF, Poor, WL, and WL (10+)). This indicates that these variables exerted a significant influence on the first component. Component 2, on the other hand, demonstrated a dominant effect of climatic factors, specifically the average temperature. The third component was found to be associated with alcohol and tobacco consumption, implying that smoking and alcohol habits may play a crucial role in the transmission of the virus. The fourth

component pertained to health insurance, while the fifth component represented the total number of health centers, encompassing sub-centers, Primary Health Centers (PHCs), and Community Health Centers (CHCs).

3.2.2. Comparison of Spatial Models

The data for all the independent variables taken in this study covers a time period of 2019 prior to the pandemic, with a constant value for both waves. To investigate the effects of these variables on COVID-19 cases and associated deaths, analyzed the data as a whole duration. This study used four spatial model (SLM, SEM, GWR, MGWR) to measure the impact of independent variables on the spatial dispersion of COVID-19 cases and deaths.

Results of the spatial modelling show that all five components have a statistically significant influence on the impact of COVID-19 in Indian districts. However, the level of influence varies among the models, and these associations remain consistent across all models, regardless of the spatial dependencies included. Based on the AIC score and R-square value, the MGWR model provides the best fit for the entire period (see table 5). The results of the MGWR model offer a reasonable explanation for the variation in COVID-19 confirmed cases and deaths across Indian districts. The coefficient estimates for all factors in the MGWR model will be interpreted in the following section.

Table 5: Comparison criteria for spatial regression model.

		SLM	SEM	GWR	MGWR
Tcc	R^2	0.40	0.41	0.52	0.57
	AIC	1490	1495	1380	1359
Deaths	R^2	0.42	0.44	0.60	0.65
	AIC	1485	1432	1395	1222

3.2.3. MGWR model summary

The table 6 presents the summary statistics of the MGWR model, providing a comprehensive overview of the model's results and key statistical measures. The MGWR, a local regression model that allows for spatially varying coefficients, was used to examine the percentage of variation explained by different spatial units. Analysis showed significant spatial variability in the R-squared approximation, with generally large values observed across districts, indicating that the regression analysis successfully explained most of the variation in the dependent variable. However, low R-squared values in some regions suggest that unexpected differences exist that are not fully captured by the independent variables in the regression model. To explore the relationship between independent variables and TCC and the total number of deaths due to COVID-19, we mapped the R-squared values for each of the 626 districts in India. With reference to the table 5 & figure 5 and 6, the local R-squared for TCC varied between 0.75 and 0.010, indicating that some local models fit better than others. A spatial regularity was observed in the distribution of R-squared values, with the most significant proportion of variation explained by the five components in more prominent parts of Madhya Pradesh- Sagar, Jabalpur, Narshimpur, Raisen districts have value of R^2 is 0.75, and districts like Porbandar, Junagarh, Somnath

districts of Gujarat have value of R^2 is 0.74. Further, districts of Maharashtra, Gujarat, Uttar Pradesh, West Bengal, and some district of Chhattisgarh demonstrating 55% to 65% of the variation in the data. The high explanatory power of independent variables was found in these districts. A notable interaction was discovered between the independent variables adopted in the study and TCC across numerous districts in Karnataka, including Vijayapura, Belagavi, and 13 others, Jharkhand, with Dhanbad, Sahibganj, and 8 other districts, Bihar, and areas nearby Maharashtra, Madhya Pradesh, Gujarat, and West Bengal. These variables accounted for 45-65% of the variance in TCC across these districts.

Table 6: Summary statistics for MGWR parameter estimates

	Variable	*Bandwidth*	*Mean*	*STD*	*Max*	*Median*	*Min*
TCC	Intercept	625	-0.096	0.009	-0.113	-0.098	-0.079
	PC1	43	0.31	0.452	-0.197	0.19	3.22
	PC2	380	0.108	0.151	-0.19	0.141	0.258
	PC3	625	-0.021	0.006	-0.029	-0.02	-0.012
	PC4	89	0.132	0.211	-0.263	0.077	1.041
	PC5	545	0.03	0.066	-0.008	-0.002	0.214
	Variable	*Bandwidth*	*Mean*	*STD*	*Max*	*Median*	*Min*
DEATHS	Intercept	625	-0.086	0.006	-0.097	-0.087	-0.074
	PC1	43	0.305	0.51	-0.16	0.167	4.224
	PC2	380	0.072	0.128	-0.196	0.133	0.188
	PC3	625	-0.047	0.039	-0.143	-0.024	-0.004
	PC4	89	0.19	0.294	-0.073	0.105	1.468
	PC5	545	0.021	0.042	-0.053	0.005	0.118

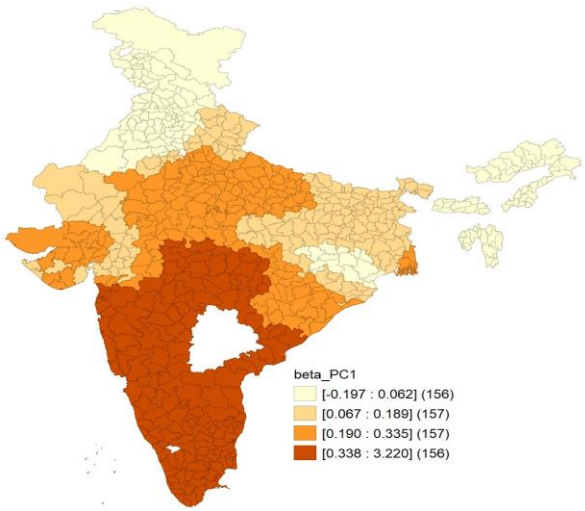
TCC- Total number of confirmed cases, Deaths- total number of deaths, Mean, Median, Max, Min- mean, median, maximum and minimum value for the coefficient estimate, STD-standard deviation, Bandwidth-number of neighbours used for the estimation

The MGWR model highlighted a positive relationship between TCC and 1, 2, 4 and 5 components and a negative association with component 3 in these regions, with significant R-squared values. In these regions, the average coefficient value for component 1, which represents age, educational level, and housing condition, was 0.310. Component 2, which indicates average temperature, was associated with a 0.10% increase in TCC per unit increase in these components in a specific district. This component exhibited higher coefficient values, particularly in districts with high average temperatures. Components 4 and 5, which reflect the dominant effects of health insurance and health centers, showed a positive association with TCC. On the other hand, component 3, representing alcohol and tobacco consumption, associated with TCC. This suggests that districts with higher alcohol and tobacco consumption highly affected by COVID-19. The variables utilized in this study were unable to explain the variation in the districts of the northeastern region and some districts of Punjab (such as Bhatinda, Faridkot, Moga, and others with $R^2 < 0.05$), Haryana (Sirsa, Panchkula with $R^2 < 0.17$), Himachal Pradesh, JK, and Ladhak, where the R-squared values ranged from 0.02 to 0.35. Hence, the reason behind the anomalous ranges of coefficient estimates in these regions becomes apparent.

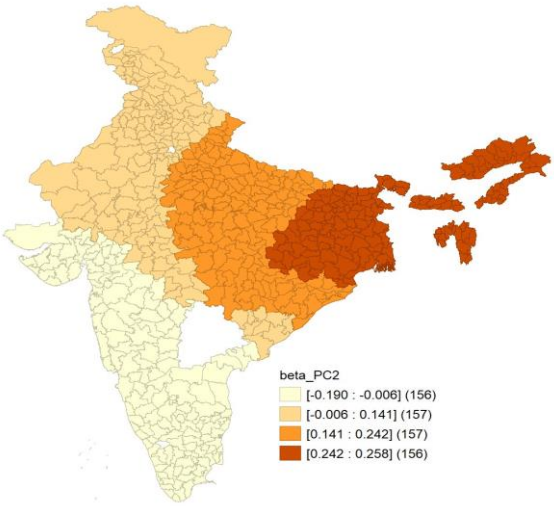
The MGWR model was used to analyze the relationship between 16 independent

variables and the total number of COVID-19 cases and deaths in various districts of India. The results showed that the five components which is combination of these independent variables explained on an average 57% of the variance for total cases and 65% for deaths. The model variances were better captured for fatalities, with significantly increased explanatory capacity shown across the models.
The most significant proportion of the variation considered local R-square values for the total number of deaths was explained by the 5 components in almost same districts of Madhya Pradesh, Maharashtra, Gujarat, Uttar Pradesh, West Bengal, Chhattisgarh where the R-square is high for the total number of confirmed cases. In these districts, these factors explain the variation in the total number of deaths associated with COVID-19 were 70% to 87%. Some fewer interaction effects were found between these components and the total number of deaths in the regions of Jammu Kashmir, Ladakh, Uttarakhand, Bihar, and Orissa, where factors used in this analysis can explain the 50% to 70% variation in the total number of deaths.

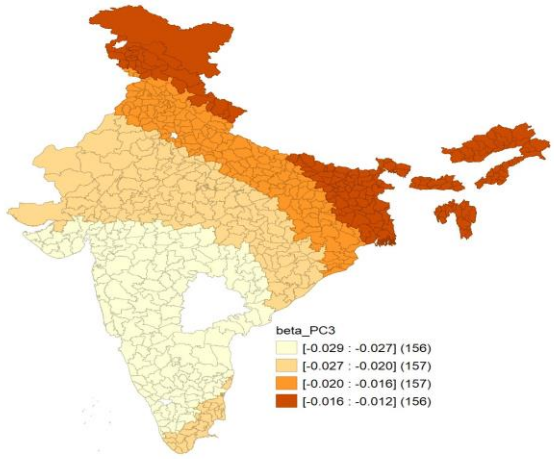
Again, the MGWR Model reveals a positive relationship between the total number of deaths and components 1, 2, 4 and 5, while displaying a negative association with factor 3 in those regions, with a significant R-squared value. In these regions, the average coefficient value for component 1 is 0.305, indicating a dominant effect of age, educational level, and housing condition. It suggests that districts with favourable housing conditions, well-educated women, and a higher proportion of adults were most affected by COVID-19. Furthermore, a unit increase in component 2 is associated with a 0.072% increase in the total number of deaths. Components 4 and 5 exhibit coefficient ranges of (-0.073, 1.468) and (-0.053, 0.118) respectively. However, their mean coefficient values are 0.195 and 0.021 respectively, suggesting that, on average, these components have a positive association with the total number of deaths. Similar to TCC, the total number of deaths in a particular district associated with tobacco and alcohol consumption. Nevertheless, the variables employed in this study are unable to account for the variation in the total number of deaths associated with COVID-19 in the districts of the northeastern region, Punjab, Haryana, and Karnataka. In these regions, the R-squared values are considerably low, ranging from 0.02 to 0.22.



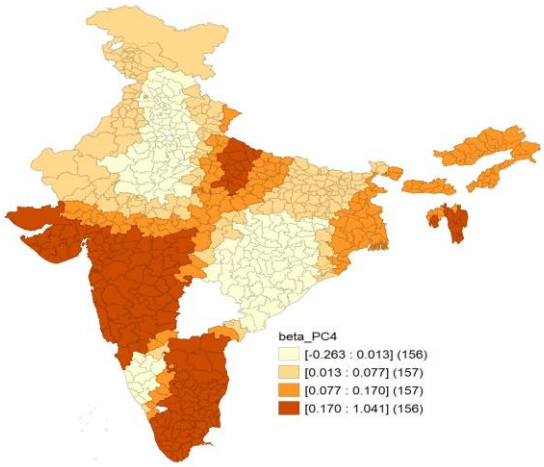
(a)



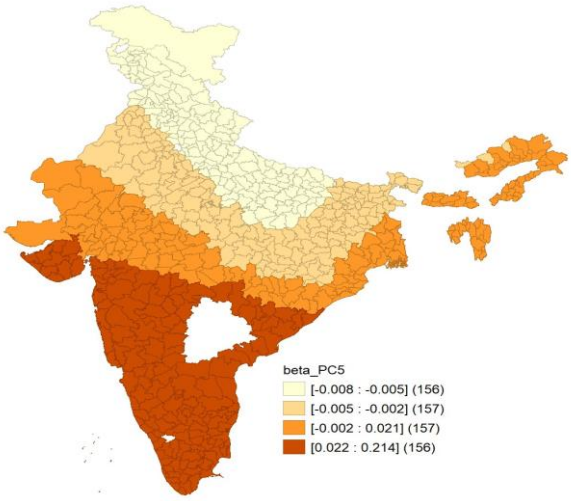
(b)



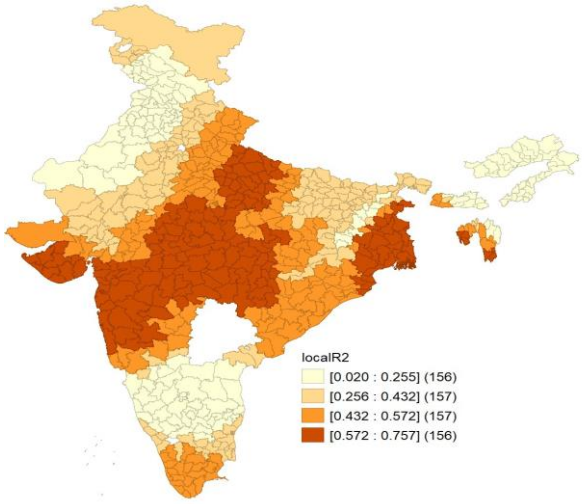
(c)



(d)



(e)



(f)

Figure 5: Spatial distribution of MGWR coefficient estimates for (a) Component first, (b) Component second, (c) Component third, (d) Component forth, (e) Component fifth, (f) Local R-squared value for the total confirmed cases of COVID-19 in 626 districts of India

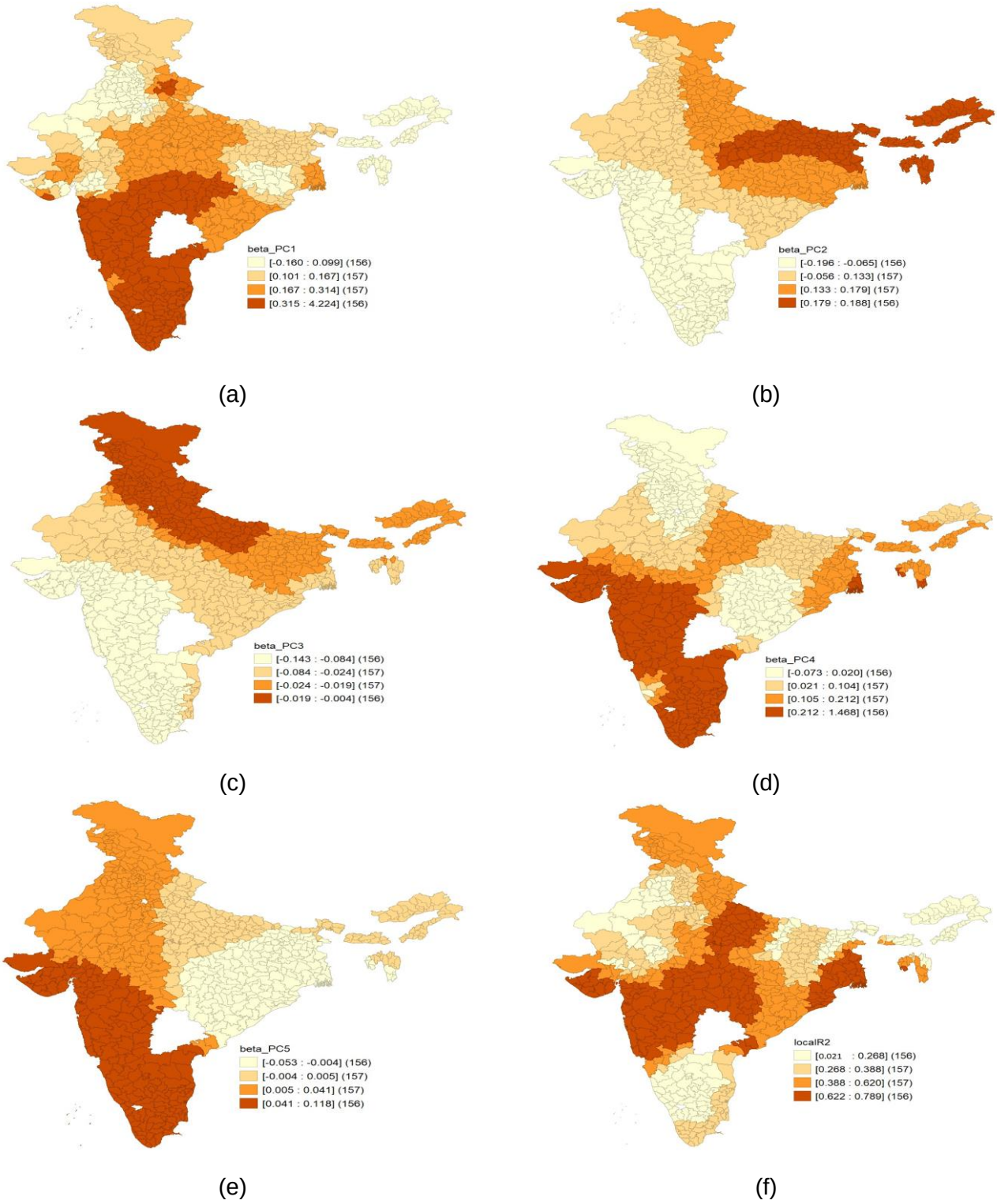


Figure 6: Spatial distribution of MGWR coefficient estimates for (a) Component first, (b) Component second, (c) Component third, (d) Component forth, (e) Component fifth, (f) Local R-squared value for the total number of deaths due to COVID-19 in 626 districts of India

4. Discussion

The current research implemented spatial analysis techniques to analyze the spatial distribution and clustering of COVID-19 in Indian districts. The data indicated a significant spatial heterogeneity in the distribution of COVID-19 across the country, with clusters of cases and deaths found to be almost identical for both waves with high intensity. The main reason for the lack of change in hotspots from the first to the second wave is attributed to the need to identify and monitor hotspots in the first wave properly. Further, the resurgence of cases has been linked to mass gatherings and non-adherence to safety protocols such as wearing masks, social distancing, and handwashing. Significant clustering of COVID-19 cases was identified in specific districts, including Maharashtra, Kerala, Andhra Pradesh, West Bengal, and Karnataka, forming clusters characterized by high numbers of COVID-19 cases and deaths. Conversely, districts in the northern and southern regions formed clusters with low COVID-19 cases and deaths. These findings imply that the risk of infection was not independent across districts. The observed spatial autocorrelation suggests that the disease may spread from high-risk districts to neighbouring areas, underscoring the importance of coordinated efforts to control the spread of the disease across all districts. The findings of this study suggest that proper identification and monitoring of hotspots in the first wave could have enabled more effective management of COVID-19 cases in the second wave.

Spatial models have demonstrated their usefulness as tools for comprehending and examining pandemic behaviour. Nevertheless, the issue of multicollinearity often poses a challenge for these models. In the present study, it was observed that the independent variables utilized to identify risk factors exhibited a considerable degree of collinearity. To address this issue, PCA was initially applied, facilitated the identification of five components. These components are derived through linear combinations of the independent variables and possess no correlation with one another. Further these principal components served as crucial inputs for the spatial models which helps in deeper exploration of the relationship between the variables and the spatial pattern of the phenomena under investigation.

In the existing literature, notable studies (Dutta et al., 2021; Sridhar, 2023) have significantly advanced our understanding of various facets related to India. However, it is noteworthy that these studies often fall short in their examination of district-level dynamics. However, these studies often lack a granular examination of district-level dynamics, which is crucial for capturing localized variations and tailoring interventions accordingly. District-level analyses are vital, particularly in a country as diverse and multifaceted as India, where regional variations can be substantial. Furthermore, an often-neglected concern in these analyses is the presence of multicollinearity, a statistical challenge that can compromise the accuracy of findings. Given the intricate interplay of factors contributing to the current situation, this study endeavours to bridge these gaps by incorporating a nuanced district-level analysis while simultaneously addressing the complexities introduced by multicollinearity. This holistic approach aims to provide a more robust and applicable understanding of the factors influencing the scenario under investigation.

In this research, MGWR model outperformed the other implemented spatial models in this study. By incorporating spatially varying coefficient, MGWR captured the local variations and heterogeneity in the relationship between the dependent and

independent variables. This study identified crucial independent variables that strongly influence the COVID-19 cases and deaths. The findings demonstrated that factors such as age structure, educational level among women, housing conditions, climatic conditions, alcohol and tobacco consumption, number of health centres, and the proportion of people with health insurance were significantly associated with COVID-19 cases and deaths. Additionally, the findings of the MGWR model demonstrated a positive relationship between high temperatures and the spread of the COVID-19 virus. This relationship is supported by epidemiological evidence indicating that an increase in ambient temperature can result in a higher transmission rate. The virus can endure in the air longer at higher temperatures and be more easily transmitted through droplets. Additionally, greater access to healthcare facilities was positively correlated with more accurate diagnosis and reporting of COVID-19 cases, which may explain the higher number of cases and deaths in these areas. Furthermore, areas with a high proportion of the population having alcohol and tobacco consumption, and high literacy rates among women were also positively associated with higher COVID-19 case. Smoking and drinking habits can weaken the immune system and make individuals more susceptible to the virus. High literacy rates among women could increase awareness of the virus and its symptoms, increase testing, more accurate diagnosis and reporting of cases, and increase transmission opportunities.

The MGWR model allows for the coefficients to vary spatially, capturing the spatial heterogeneity of the relationships between the variables. The range of coefficients provides valuable information about how the relationships between these variables change across space. The findings of the MGWR model shed light on the extent to which the identified risk factors explain the variation in COVID-19 cases and deaths across different districts. Notably, districts such as Mumbai, Chennai, Pune, Kolkata, Sagar, Jabalpur, Narshimpur, Raisen, Porbandar, Junagarh, and Somnath demonstrate a substantial proportion of variation in COVID-19 outcomes (ranging from 75 to 85 percent) that can be accounted for by these factors. This indicates the strong influence of the identified risk factors in these regions. Conversely, the variables considered in this study were insufficient in explaining the variation observed in certain districts, primarily in the northeastern region and some districts of Punjab (e.g., Bhatinda, Faridkot, Moga). Additionally, limited explanatory power was observed in regions such as Sirsa, Panchkula, and districts of Himachal Pradesh, JK, and Ladhak. These anomalous ranges of coefficient estimate in these regions suggest that other unaccounted factors may play a more significant role in shaping COVID-19 outcomes.

The overall findings suggest that addressing multicollinearity in spatial models can significantly enhance their robustness and reliability. By mitigating the impact of collinearity among independent variables, researchers can obtain more accurate and trustworthy results. Consequently, this enables the identification of high-risk districts where targeted interventions can be implemented. Measures such as rigorous testing and contact tracing, targeted lockdowns, and intensified public health messaging can be strategically deployed to effectively control and mitigate the spread of the virus in these specific areas. However, limitations of the study include its reliance on reported case counts and its focus on only two waves of the pandemic due to data

unavailability, which may not capture the full impact of the virus. Therefore, future research should address these shortcomings to develop more effective strategies for mitigating them.

5. Conclusion

This study aimed to employ spatial econometric modelling methods to enhance understanding of the spatial structures and associations among locations in India and to analyses the transmission patterns of COVID-19. By considering spatial proximity, the study assessed the impact of demographic, socioeconomic, climatic, and comorbidity on total COVID-19 cases and deaths across 640 districts in India. Additionally, this study addressed the issue of multicollinearity in spatial models through the utilization of principal component analysis. This approach successfully reduced interdependence among variables and improved the model's accuracy, allowing for the identification of key risk factors associated with the phenomenon under investigation. The results underscore the importance of dealing with multicollinearity in spatial models and offer practical implications for decision-making and policy formulation. Notably, the study revealed that household conditions, educational level of women, tobacco and alcohol consumption rates, the number of health centers, and climatic factors were influential in COVID-19 incidence. Furthermore, the study identified areas with older populations as having higher COVID-19 cases. The findings of this study can inform the development of prevention strategies and strengthen public health capacities, particularly in regions where the healthcare system may be limited. However, it is worth noting that a limitation of the analysis was the lack of district-level data on deaths beyond October 2021 in India.

References

Adekunle, I. A., Onanuga, A. T., Akinola, O. O., & Ogunbanjo, O. W. (2020). Modelling spatial variations of coronavirus disease (COVID-19) in Africa. *Science of The Total Environment*, *729*, 138998. https://doi.org/10.1016/j.scitotenv.2020.138998

Ahmed, F., Ahmed, N., Pissarides, C., & Stiglitz, J. (2020). Why inequality could spread COVID-19. *The Lancet Public Health*, *5*(5), e240. https://doi.org/10.1016/S2468-2667(20)30085-2

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Chen, B., Liang, H., Yuan, X., Hu, Y., Xu, M., Zhao, Y., Zhang, B., Tian, F., & Zhu, X. (2020). *Roles of meteorological conditions in COVID-19 transmission on a worldwide scale* [Preprint]. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.03.16.20037168

Dutta, I., Basu, T., & Das, A. (2021). Spatial analysis of COVID-19 incidence and its determinants using spatial modeling: A study on India. *Environmental Challenges*, *4*, 100096. https://doi.org/10.1016/j.envc.2021.100096

Dzul-Manzanilla, F., Correa-Morales, F., Che-Mendoza, A., Palacio-Vargas, J., Sánchez-Tejeda, G., González-Roldan, J. F., López-Gatell, H., Flores-Suárez, A. E., Gómez-Dantes, H., Coelho, G. E., Da Silva Bezerra, H. S., Pavia-Ruz, N., Lenhart, A., Manrique-Saide, P., & Vazquez-Prokopec, G. M. (2021). Identifying urban hotspots of dengue, chikungunya, and Zika transmission in Mexico to support risk stratification efforts: A spatial analysis. *The Lancet Planetary Health*, *5*(5), e277–e285. https://doi.org/10.1016/S2542-5196(21)00030-9

Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale Geographically Weighted Regression (MGWR). *Annals of the American Association of Geographers*, *107*(6), 1247–1265. https://doi.org/10.1080/24694452.2017.1352480

Gupta, A., Banerjee, S., & Das, S. (2020). Significance of geographical factors to the COVID-19 outbreak in India. *Modeling Earth Systems and Environment*, *6*(4), 2645–2653. https://doi.org/10.1007/s40808-020-00838-2

Mollalo, A., & Khodabandehloo, E. (2016). Zoonotic cutaneous leishmaniasis in northeastern Iran: A GIS-based spatio-temporal multi-criteria decision-making approach. *Epidemiology and Infection*, *144*(10), 2217–2229. https://doi.org/10.1017/S0950268816000224

Oshan, T., Li, Z., Kang, W., Wolf, L., & Fotheringham, A. (2019). mgwr: A Python Implementation of Multiscale Geographically Weighted Regression for Investigating Process Spatial Heterogeneity and Scale. *ISPRS International Journal of Geo-Information*, *8*(6), 269. https://doi.org/10.3390/ijgi8060269

Pereira, M., & Oliveira, A. M. (2020). Poverty and food insecurity may increase as the threat of COVID-19 spreads. *Public Health Nutrition*, *23*(17), 3236–3240. https://doi.org/10.1017/S1368980020003493

Pfeiffer, D. U., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., & Clements, A. C. A. (2008). *Spatial Analysis in Epidemiology*. OUP Oxford.

Puebla Neira, D., Watts, A., Seashore, J., Polychronopoulou, E., Kuo, Y.-F., & Sharma, G. (2021). Smoking and risk of COVID-19 hospitalization. *Respiratory Medicine*, *182*, 106414. https://doi.org/10.1016/j.rmed.2021.106414

Sarkar, S. K., Ekram, K. M. M., & Das, P. C. (2021). Spatial modeling of COVID-19 transmission in Bangladesh. *Spatial Information Research*, *29*(5), 715–726. https://doi.org/10.1007/s41324-021-00387-5

Shen, D., & Zhu, H. (2015). Spatially Weighted Principal Component Regression for High-Dimensional Prediction. In S. Ourselin, D. C. Alexander, C.-F. Westin, & M. J. Cardoso (Eds.), *Information Processing in Medical Imaging* (pp. 758–769). Springer International Publishing. https://doi.org/10.1007/978-3-319-19992-4_60

Sridhar, K. S. (2023). Urbanization and COVID-19 Prevalence in India. *Regional Science Policy & Practice*, *15*(3), 493–505. https://doi.org/10.1111/rsp3.12503

Tosepu, R., Gunawan, J., Effendy, D. S., Ahmad, L. O. A. I., Lestari, H., Bahar, H., & Asfian, P. (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Science of The Total Environment*, *725*, 138436. https://doi.org/10.1016/j.scitotenv.2020.138436

Wang, Q., Dong, W., Yang, K., Ren, Z., Huang, D., Zhang, P., & Wang, J. (2021). Temporal and spatial analysis of COVID-19 transmission in China and its influencing factors. *International Journal of Infectious Diseases*, *105*, 675–685. https://doi.org/10.1016/j.ijid.2021.03.014