Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 143-160

Exploring COVID-19 Spatial Patterns in Indian Districts: Ridge and Lasso Geographic Weighted Models for Spatial Heterogeneity and Multicollinearity

Megha Sharma[1] **and Shalini Chandra**[1]
[1]*Department of Mathematics and Statistics*
Banasthali Vidyapith, Rajasthan, India

Received: 22 June 2024; Revised: 28 July 2024; Accepted: 31 July 2024

Abstract

This study conduct a comprehensive spatial analysis of COVID-19 across districts in India utilizing 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 of the study is to uncover risk factors through spatial modelling while mitigating multicollinearity using the concept of LASSO and ridge regression. Employing spatial analysis, we identify COVID-19 hotspots and coldspots across districts. High-impact districts including Mumbai, Pune, Chennai, Kolkata, and Bengaluru are highlighted, along with lesser-affected districts in central and north-eastern regions. Analysis used geographical weighted regression (GWR) models, incorporating ridge and LASSO techniques to assess the impact of demographic, socioeconomic, climatic, and comorbidity factors on COVID-19 while accounting for spatial relationships. Notably, the GWR with LASSO (GWL) outperforms the other models, with lower RMSE and a notably higher R^2 value. This study reveal significant risk factors such as sanitation facilities, healthcare amenities, women's education, tobacco/alcohol usage, urban population and density, comorbidity, as well as climatic conditions. The GWL model's localized coefficients offer valuable insights into predictor relationships within each spatial unit.

Key words: COVID-19; Geographic weighted model; LASSO regression; Spatial association; Ridge regression.

1. Introduction

The COVID-19 pandemic began in Wuhan, China, in December 2019, caused by the SARS-CoV-2 virus (Li *et al.*, 2020). The COVID-19 pandemic has had a profound impact on individuals' lives, the global economy, and public health. India has been hit particularly hard, suffering economic disruption, unemployment, and a decline in GDP due to COVID-19. The country's healthcare system struggled with resource shortages, limited

Corresponding Author: Megha Sharma
Email: meghasharma15aug@gmail.com

hospital space, and personnel shortages (Sridhar, 2023, Dutta *et al.*, 2021). The pandemic has also triggered social and psychological issues, including increased domestic violence, mental health challenges, and gender inequality (Sardar *et al.*, 2020). Numerous previous studies have identified that social inequalities can facilitate the spread of diseases (Ahmed *et al.*, 2020). Poor living conditions (Pereira and Oliveira, 2020), population density (Rocklöv and Sjödin, 2020), inadequate access to healthcare, and a large proportion 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. Temperature has also been associated with COVID-19 severity, with similar findings in China (Chen *et al.*, 2020), Indonesia (Tosepu *et al.*, 2020), Turkey (Chung *et al.*, 2021), and the USA (Bashir *et al.*, 2020). Additional risk factors like the prevalence of slums within cities (Sridhar, 2023), smoking habits, and many more contribute to an increased risk of transmission and disparities in access to prevention and treatment measures.

Spatial models have emerged as valuable tools for determining the relationships between the spread of infectious diseases and associated risk factors, incorporating the spatial dimension. Spatial methods are employed to model particular variables at diverse geographical locations, allowing us to address the diversity caused by regional differences (known as spatial heterogeneity) within the data. One effective method for identifying spatial heterogeneity is the Geographically Weighted Regression (GWR) model, which is highly effective in accurately estimating parameters when analyzing COVID-19 data (Sarkar *et al.*, 2021, Ramírez-Aldana *et al.*, 2020, Appiah-Otoo and Kursah, 2022, Adekunle *et al.*, 2020). The GWR model helps illustrate how the association between independent and dependent variables varies across distinct locations within the study area. However, a challenge arises when the risk factors examined within each local model exhibit linear relationships, which is referred to as local multicollinearity. This multicollinearity issue obstructs the precision of parameter estimates and makes it difficult to distinguish the individual effects of these variables.

In the context of addressing the challenge posed by multicollinearity in data, various alternative methodologies have arisen as effective solutions. One such prominent technique is ridge regression, initially proposed by Hoerl and Kennard in 1970 (Hoerl and Kennard, 1970), which has become widely adopted for mitigating the issues associated with multicollinearity. This shrinkage technique incorporates penalty terms into the regression framework to shrink the coefficients, resulting in more stable parameter estimates and mitigate the effect of multicollinearity. Ridge regression introduces a positive bias into the parameter estimation process, effectively guiding the coefficients towards zero. Although this approach yeilds baised results, it reduces variance. Recognizing the potential benefits of combining different methodologies, researchers have explored various approaches, such as combining ridge regression with the Liu estimator (Kejian, 1993) or integrating ridge regression with principal component regression (Baye and Parker, 1984, Chandra and Sarkar, 2016), among others. Additionally, in 1996, Tibshirani introduced a novel technique that has gained extensive attention. The technique combines the advantages of ridge regression with variable selection method, known as the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996). LASSO leverages the LARS algorithm (Least Angle Regression) to shrink estimated coefficients towards zero and selectively sets less significant variables to precisely zero. The resulting model is notably interpretable, retaining only the most meaningful predictors relevant to the outcome variable.

Similar to their role in classical regression models, Ridge and LASSO techniques serve to mitigate the impact of multicollinearity in spatial context. Likewise, in the context of GWR model, tackling local multicollinearity is achievable through the incorporation of ridge regression and the LASSO method. Specifically, the utilization of ridge regression within the GWR framework is referred as GWRR, while the integration of the LASSO method with GWR as GWL (Wheeler, 2007). In this study, GWRR and GWL models were applied to investigate spatial heterogeneity and address multicollinearity concerns during the analysis of the COVID-19 pandemic across 626 districts in India.

2. Data

2.1. Data collection

This study included districts from all states and union territories in India, except six states: Assam, Delhi, Goa, Manipur, Telangana, and Sikkim, with no available COVID-19 updates at the district level in the state bulletin. This study took into account district boundaries as of 2019. We extracted district-level data on daily confirmed cases of COVID-19 and associated deaths in India 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 availability of data until October 2021. 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 into our study, and the sources from which they were obtained.

The data were sourced from various sources, with the primary contributor being National Family Health Survey (NFHS-5). It is initiated by the Ministry of Health & Family Welfare (MoHFW), Government of India, plays a vital role in assessing health conditions in India. This extensive health survey is periodically conducted nationwide, offering health-related indicators at the district, state, and national levels. NFHS-5 was conducted in India during the time period of 2019-2021.

2.2. Data preparations and cleaning

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. This research studied the cumulative confirmed cases (CCC) and cumulative deaths ($deaths$) during the first and second waves at the district level in India. To facilitate the analysis, initial data preparation with a data cleaning procedure aimed at addressing issues like incomplete and duplicate entries by cross-referencing data from various sources. Given the multi-sourced nature of the data, a pivotal step involves data merging, wherein information from diverse origins is consolidated to establish a unified reference point. As the data on risk factors and COVID-19 incidences originates from distinct sources, it is imperative to standardize the data before any analysis can take place. Standardization, in this context, entails bringing different variables onto a common

Table 1: Lists of independent variables, their descriptions and justification and the sources

Abre-viation	Indicators	Assumptions/ Justifications	Data sources
V1	Population below age 15 years	Older population have higher risk of death after infected.	National Family Health Survey (NFHS-5) (2019-21) (District factsheet)
V2	Population living in households with electricity	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.	
V3	Population living in households that use an improved sanitation facility		
V4	Households using clean fuel for cooking		
V5	Households with any usual member covered under a health insurance	Accessible healthcare systems, affordability, capacity, and health security are vital for managing epidemics and promoting treatment-seeking.	
V6	Women who are literate	Women's literacy empowers them with knowledge, enabling them to understand COVID-19 prevention, access reliable information, and make informed decisions.	
V7	Educated women with 10+ years of schooling		
V8	Proportion of women undernourished	Undernourished and obesity weakened immunity and elevate risk due to underlying health problems when facing COVID-19.	
V9	Proportion of women obese		
V10	Tobacco use among those 15+	Smoking or tobacco or any kind of alcohol being exposed in any form can reduce the risk of COVID-19 infection (WHO 2020).	
V11	Alcohol use among those 15+		
V12	Population 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
V13	Proportion of urban population		
V14	Health Center [Sub center +PHCs+ CHCs]	Higher population per healthcare institution lower resilience in dealing with COVID-19.	Rural Health statistics
V15	Hypertension among Adults	Blood Sugar Level and Hypertension among Adults (age 15+) may regulate the severity of COVID-19 cases.	(NFHS-5)
V16	Adults' blood sugar levels (Age 15+)		
V17	Average temprature	The severity of COVID-19 associated with temperature and relative humidity	NASA open data portal
V18	Relative Humidity		
V19	Proportion of poor population	Studies have shown that areas with high poverty rates tend to have higher rates of COVID-19 infections.	Global Data Lab

scale to enable comparisons across variable types. These steps were undertaken prior to the transformation of the data into district-level counts and its merging with India's district administrative boundary shape file from the DIV-GIS database using ArcGIS Desktop 10.7.

2.3. Data description

This study involves 19 dependent variables and two independent variables (*CCC* and *deaths*), with data collected from diverse sources across 626 districts in India, resulting in around 13,000 observations—an extensive dataset for analysis. Emphasizing spatial analysis as the foundation, we prioritize reviewing the data before applying statistical methods.

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

According to the data, higher population density is observed 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 display 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 reveal pronounced water supply challenges in numerous districts of Odisha, Madhya Pradesh, and Rajasthan. Noteworthy is that 12 out of 14 districts in Kerala and 21 out of 30 districts in Tamil Nadu are in the highest quartile (>27%) for the proportion of women grappling with obesity. 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 highlight certain 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. Alcohol and tobacco consumption is notably high in districts of northeastern states, Punjab, Goa, and select districts in Rajasthan.

According to the National Family Health Survey (NFHS-5), about 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%.

2.4. Visualization and exploration

By using visualization techniques, patterns and discrepancies in the data are 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. Such maps have been included in the study to present the spatial distribution of COVID cases and deaths in further sections. Exploration of spatial data includes cluster analysis to identify whether observed spatial patterns are random, using either nonspecific (global) or specific (local) techniques. Moran's I statistic, a global technique, is employed to ascertain cluster presence across the entire study area. Moran's I computes global spatial autocorrelation among observations and ranges from -1 to 1. Negative values indicate dispersion (clustering of dissimilar values), positive values indicate clustering (clustering of similar values), and values near zero suggest absolute spatial randomness, implying no autocorrelation. However, because Moran's I statistic is incapable of providing precise information on cluster locations, the LISA (local indicators of spatial association) tool was utilized to calculate local spatial autocorrelation. This method describes significant correlations at specific locations as local spatial clusters (hot spots) or correlations between observations and neighboring observations Anselin (1995). The next section is about the models and estimators considered in this study.

3. The models and estimators

3.1. Geographical weighted regression (GWR)

The GWR model estimated local interactions between the dependent and independent variables by fitting a regression model to each feature (spatial unit) in the dataset (Brunsdon *et al.* (1998)). The GWR model for each feature is

$$y_i = \beta_{i0} + \sum_{j=1}^{m} X_{ij}\beta_{ij} + \epsilon_i, i = 1, 2, ..., n. \tag{1}$$

where y_i represents the dependent variable at a specific location i, β_{i0} stands for the intercept parameter at that same location i, β_{ij} symbolizes the local regression coefficient pertaining to the j^{th} explanatory variable at location i, X_{ij} signifies the value of the j^{th} explanatory variable at location i, and ϵ_i corresponds to the random error observed at location i. The parameters estimates for each independent variable at i^{th} location is given by

$$\hat{\beta(i)} = (X^T W(i) X)^{-1} X^T W(i) y \tag{2}$$

where $\hat{\beta(i)}$ is $m \times 1$ vector of parameter estimates, $W(i)$ is spatial weight matrix calculated by the exponential kernel function which is defined as

$$w_k(i) = \begin{cases} \left[1 - \left(\frac{d_{ik}}{bw}\right)^2\right]^2, & \text{if } k \in \{N_i\} \\ 0, & \text{if } k \notin \{N_i\} \end{cases} \tag{3}$$

where d_{ik} is the distance between feature location i & k with bandwidth bw derived from the Euclidean distance between observation locations and neighboring points, this measure

ensures that the region remains influenced by proximate neighbors within this radius. The set N_i includes observations within this N^{th} nearest neighbor distance. Weights are zero for observations beyond this range, except for observation i which gets a weight of 1. Kernel function assigns higher weights to observations that are closer to the calibration location i. To fit the GWR model, the kernel bandwidth is estimated through cross-validation (CV) using all feature locations, followed by weight calculation using (3). CV function is outlined as

$$CV(bw) = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(bw)]^2 \tag{4}$$

where, $\hat{y}_{\neq i}(bw)$ is the estimated value of y achieved by excluding the data point at the i^{th} location during prediction. The bandwidth bw will be derived through an iterative procedure aiming to minimize the CV score.

3.2. Addressing multicollinearity: diagnosis and remediation

Collinearity's presence among independent variables can diminish the precision of coefficients (Wheeler and Tiefelsdorf (2005)). There are valuable diagnostic tools designed to uncover collinearity issues that might disrupt the interpretation of estimated regression coefficients. These diagnostic methods are derived from conventional regression techniques. Approaches for identifying collinearity among independent variables comprise metrics like variance inflation factors (VIF) and condition indices. Moreover, Ridge regression and LASSO are frequently employed methods for mitigating the multicollinearity.

Ridge regression

Ridge regression was uniquely formulated to alleviate the impacts of collinearity through the imposition of penalties on the magnitudes of regression coefficients. This strategy diminishes the impact of variables with comparatively low variance within the model. The parameter for ridge regression is determined by minimizing the sum of squared errors, introducing constraints that compel coefficients to approach zero (Hoerl and Kennard, 1970). More precisely, the ridge estimator coefficient is derived by minimizing the equation

$$\hat{\beta}_R = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_i \right)^2 \tag{5}$$

with $\sum_{j=1}^{m} (\beta_i)^2 \leq \rho$, where ρ is a control shrinkage amount. Then parameter estimates is obtained by

$$\hat{\beta(i)}_R = (X^T X + CI)^{-1} X^T y \tag{6}$$

where I is an identity matrices and C represents positive coefficient bias.

Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is the regularization and penalization technique which shrinks the regression coefficients towards zero, also puts least significant variable coefficients to zero. This leads to a simplified and interpretative model, retaining only the significant predictors for the outcome variable (Tibshirani, 1996). The coefficients of Lasso parameters cannot be

directly calculated through closed-form equations, unlike Ridge regression. Instead, they are determined using quadratic programming techniques. LASSO is defined as follows as

$$\hat{\beta} = argmin \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 . \tag{7}$$

where $\sum_{i=j}^{p} |\hat{\beta}_j| \leq tp$ (threshold parameter). It is established that tp is a parameter governing the level of shrinkage in LASSO coefficient estimation, where $tp \geq 0$.

3.3. Geographic Weighted Ridge Regression (GWRR)

GWRR is a modified technique within the domain of spatial regression, in which GWR model combine with the ridge regression (Wheeler, 2009). Estimator of parameters of GWRR model at i^{th} location is obtained by

$$\hat{\beta(i)}_{GWRR} = (X^T W(i) X + CI)^{-1} X^T W(i) y. \tag{8}$$

The process involves predicting bandwidth values to form a weighted matrix, minimizing bias using CV, and iteratively determining the coefficient value C for each bandwidth. These results are then applied to estimate spatial model with ridge regression coefficients.

3.4. Geographically Weighted LASSO (GWL)

LASSO's application within a GWR model, later recognized as Geographically Weighted LASSO (GWL), addresses spatial variations and local multicollinearity. GWL offers unbiased coefficient estimates and enhances prediction accuracy (Wheeler, 2009). LASSO parameter estimation in GWL is executed concurrently, relying on a pre-established kernel bandwidth. During the GWL parameter estimation process, the shrinkage (s) value is determined prior to the final LASSO solution. Shrinkage parameter estimation in GWL's LASSO model is achieved through cross-validation (CV), resulting in a distinct shrinkage parameter for each geographical location.

4. Model selection criteria

Coefficient of determination (R^2) and root mean square error (RMSE) were used to compare the performances of various models. R^2 measures the goodness of fit; its values range from 0 to 1. Furthermore RMSE calculated how closely predicted values align with actual observations by measuring the average error magnitude. The model with lower RMSE value and higher value of R^2 better fits the observed data. In this study, analysis have been performed in R software version 4.3.1 using various packages such as sp, spgwr, spdep, gwrr, and spatstat.

5. Empirical findings

5.1. Visualization and 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.

These maps use various colors to depict the intensity of variables of interest in geographic region. Referring to Figure 1, the districts that exhibited the highest numbers of confirmed COVID-19 cases and deaths were Bangaluru, Mysuru, Belagavi, and 13 other districts in Karnataka. 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 had the fewest confirmed cases and deaths in both waves. 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=0.0001[$<$ 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-2) and only a few districts fell into the high-low and low-high clusters.

5.2. Spatial modelling

The dataset encompassing all independent variables used in this study exhibits consistent values across both waves of COVID-19. With the aim of exploring the influence of these variables on the occurrences of COVID-19 cases and related fatalities, a comprehensive approach was adopted by examining the entire temporal span. The outcomes of the Global Moran's I test [Value of Global Morna's I= 0.42 for *CCC* and 0.45 for *deaths*] and the Breusch-Pagan test with p-value = 0.0001 $<$ 0.05 indicate that the data employed in this study exhibit noteworthy spatial heterogeneity.

5.2.1. Local multicollinearity

The presence of multicollinearity among independent variables can be ascertained by examining the VIF values of local observations and the condition index specific to that particular location. The summary of VIF values and condition indices is presented in the Table 2 as the GWR model incorporates all these independent variables to predict total confirmed cases and total number of deaths.

Upon referencing the table 2, it becomes evident that numerous locations exhibit VIF values and condition indices exceeding 30, indicating a significant level of concern regard-

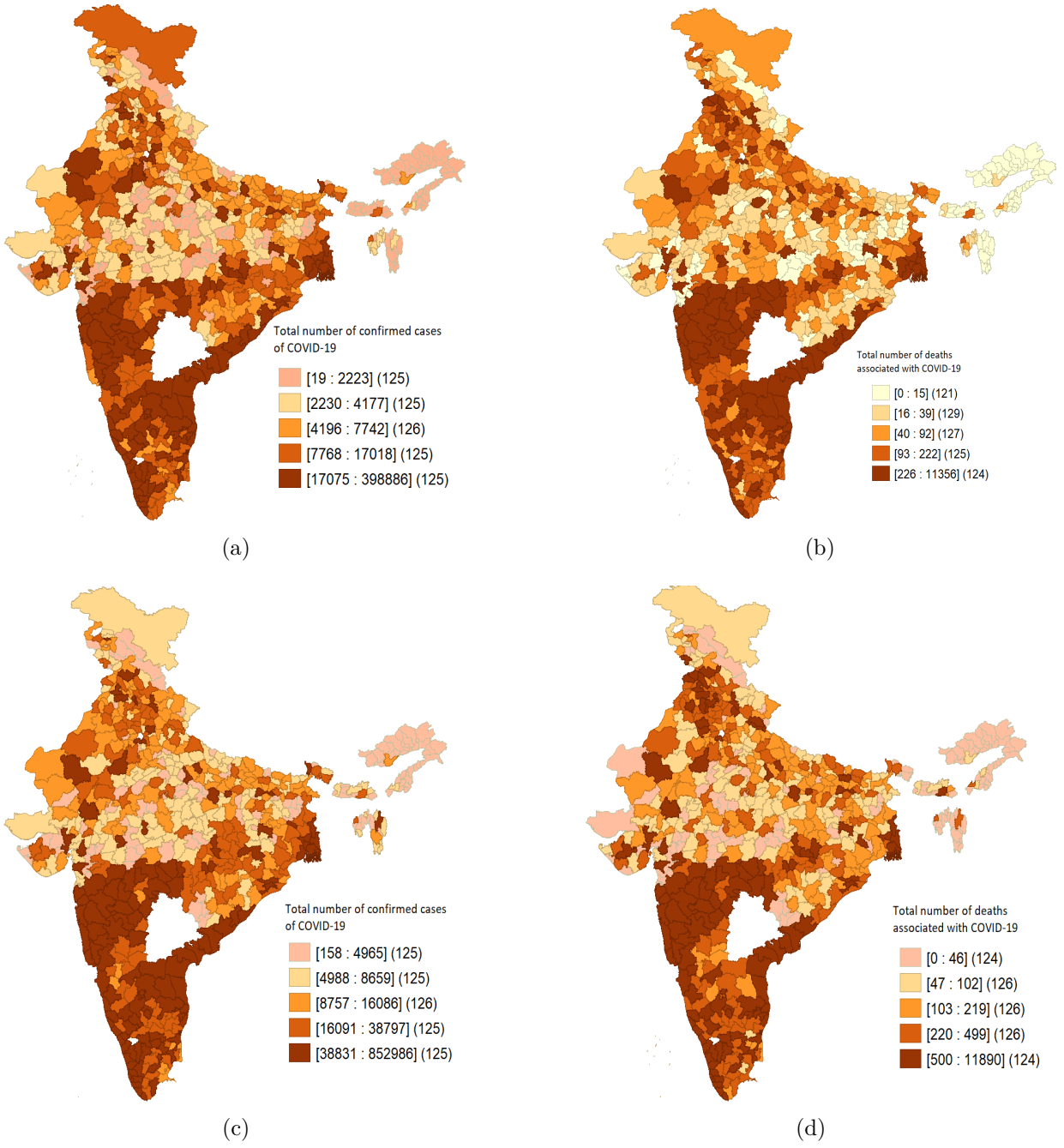

Figure 1: Quantitative spatial distribution of *Cumulative confirmed cases* (a, c) and *Total deaths* (b, d) in 1^{st} wave and 2^{nd} wave respectively in Indian districts

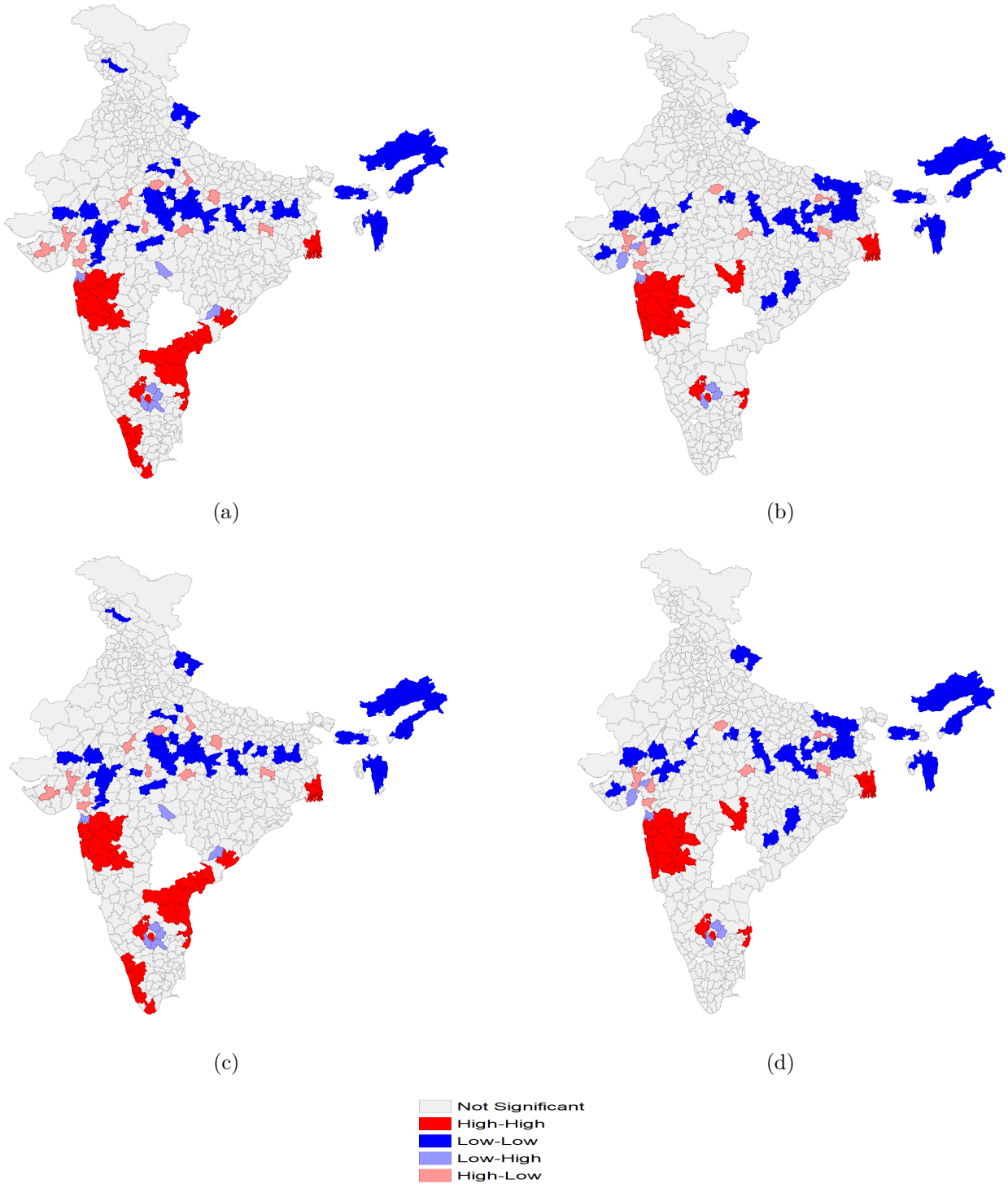

Figure 2: LISA clusters of *Cumulative confirmed cases* (a, c) and *Total deaths* (b, d) associated with COVID-19 in 1^{st} wave and 2^{nd} wave respectively in Indian districts

Table 2: Summary table of VIF and condition index at local level

	CCC			*deaths*		
VIF	Min.	Mean	Max.	min.	Mean	Max.
V1	3.049	3.23	3.416	3.037	3.954	4.955
V2	1.525	1.541	1.555	1.493	1.682	1.889
V3	2.516	2.657	2.87	2.208	3.056	5.138
V4	3.3	3.446	3.547	3.161	3.964	5.407
V5	1.339	1.419	1.452	1.327	1.828	2.244
V6	6.164	6.292	6.47	6.189	8.00	10.517
V7	5.055	5.194	5.462	4.759	6.672	9.71
V8	26.09	28.49	33.56	16.99	24.53	46.43
V9	19.36	20.12	22.43	14.28	21.31	31.89
V10	1.766	1.813	1.898	1.643	2.135	2.71
V11	1.688	1.814	2.155	1.18	1.877	3.338
V12	1.321	1.326	1.336	1.303	1.408	1.527
V13	2.509	2.551	2.593	2.302	2.63	3.096
V14	1.061	1.067	1.076	1.086	1.166	1.403
V15	25.89	28.06	32.71	17.13	23.41	42.63
V16	21.79	23.44	25.01	16.24	23.5	34.16
V17	1.829	2.008	2.336	1.275	3.074	7.281
V18	2.442	2.561	2.75	1.886	3.372	5.957
V19	2.546	2.71	2.906	1.876	3.298	4.901
CI	18.15	19.96	24.26	16.26	22.26	40.51

ing multicollinearity at particular location. This degree of multicollinearity contributes to heightened variability in the coefficient parameters, leading to less stable results. Furthermore, the existence of multicollinearity gives rise to an unstable model, a fact that becomes apparent through the modification of the classical GWR model. This modification involves the integration of multicollinearity mitigation techniques such as LASSO and ridge regression. The enhanced GWR model's effectiveness can be observed in the Table 3, where a comparison is made between the GWR model, the GWRR (Geographically Weighted Ridge Regression) model, and the GWL (Geographically Weighted LASSO Regression) model.

Table 3: Comparison table of modified and unmodified GWR model

		GWR	GWRR	GWL
CCC	$RMSE$	0.7921	0.5322	0.4630
	R^2	0.4558	0.7162	0.8371
	bw	4.9686	0.7724	0.0361
deaths	$RMSE$	0.6978	0.4597	0.4398
	R^2	0.534	0.8214	0.8625
	bw	3.1805	0.6197	0.3012

Upon evaluating the RMSE scores and R-squared values, it becomes apparent that the GWL model delivers the most accurate fit across the entire duration (refer to the Table 3). The GWL model is capable of explaining an average of 83% of the variation in cumulative

COVID-19 cases and 86% of the variation in COVID-19-related deaths across all districts of India, taking care of the challenge posed by multicollinearity. Furthermore, the GWL model provides a coherent interpretation for the disparities in confirmed COVID-19 cases and associated deaths among Indian districts. A comprehensive summary of coefficient estimates for all independent variables within the GWL model will be presented in the subsequent section.

5.2.2. GWL model summary

In GWL modelling, similar to how LASSO works, the importance of coefficients gradually decreases until they become zero due to shrinkage. When a coefficient reaches zero, it loses its influence on the outcomes of the model. Through an iteration process driven by cross-validation, GWL yields a bandwidth value of 0.09 for *CCC* and 0.89 for deaths. This bandwidth parameter, along with the associated shrinkage value, contributes to the delineation of GWL's specific parameters, all of which are detailed in the Table-4.

Referring to the summarized Table-4, it becomes evident that certain independent variables - such as Population age, households with electricity, the percentage of women who are obese, and relative humidity - possess either a zero or near-zero mean coefficient value across all regions in the context of modelling total confirmed cases. This observation signifies that these variables exert negligible influence on the incidence of COVID-19 cases. Similarly, when it comes to predicting the number of deaths, both relative humidity and household electrification also demonstrate insignificant effects. However, in contrast to COVID-19 cases, the prediction of COVID-19 deaths shows a positive association with individual age, underscoring the elevated risk of mortality among the elderly population subsequent to infection.

The cumulative confirmed COVID-19 cases in an Indian district are linked positively to factors such as the availability of sanitation facilities and healthcare services, the percentage of undernourished women, tobacco and alcohol consumption, population density, urbanization, average temperature, and the education level of women. Conversely, they are negatively associated with the number of people living in poverty. However, concerning the total number of COVID-19 related deaths, there is a negative correlation with the availability of sanitation facilities and health insurance coverage.

The data presented in the Table-4 indicates that the GWL model zeroes out coefficients for various factors in different locations, resulting in varied parameter magnitudes across regions. As a result, the GWL model generates distinct models with differing coefficients for various locations. To illustrate this, we have provided the model for the two most severely impacted districts (Pune and Bengaluru) in different zones.

$$y^*(TCC_{Pune}) = 4.55 - 0.0525V3^* + 0.115V4^* + 0.24V6^* + 0.13V7^* - 0.65V8^* + 0.11V9^* + 0.33V11^* + 0.33V13^* + 0.51V16^* + 0.20V18^*$$

$$y^*(Deaths_{Pune}) = 2.33 + 0.09V1^* - 0.001V3^* + 0.0086V4^* + 0.076V6^* - 0.047V8^* + 0.032V9^* + 0.75V11^* + 0.007V13^* + 0.078V16^* + 0.0091V17^*$$

Table 4: Summary statistics for GWL parameter estimates

	CCC					*deaths*				
	Intercept	V1	V2	V3	V4	Intercept	V1	V2	V3	V4
Min.	-2.361	-0.018	-0.016	-0.025	0.000	-2.516	-0.04321	-0.027	-0.147	0.000
1st Qu	-0.150	0.000	0.000	0.000	0.011	-0.141	0.000	0.000	0.000	0.110
median	0.000	0.000	0.000	0.000	0.034	0.000	0.000	0.000	0.000	0.137
mean	0.039	0.001	0.000	0.002	0.036	0.0257	-0.003	0.000	-0.010	0.119
3rd Qu.	0.000	0.000	0.000	0.000	0.054	0.000	0.000	0.000	0.000	0.155
Max.	4.301	0.126	0.000	0.131	0.101	4.908	0.111	0.000	0.000	0.209
	V5	V6	V7	V8	V9	V5	V6	V7	V8	V9
Min.	-0.087	-0.054	0.000	-0.591	-0.514	-0.145	0.000	-0.016	-0.660	-0.413
1st Qu	0.000	0.000	0.075	0.000	0.000	0.000	0.000	0.000	0.000	0.000
median	0.000	0.000	0.104	0.000	0.000	0.000	0.036	0.016	0.000	0.000
mean	-0.002	-0.001	0.117	0.003	-0.014	-0.001	0.048	0.020	-0.015	-0.005
3rd Qu.	0.000	0.000	0.169	0.000	0.000	0.000	0.076	0.034	0.000	0.000
Max.	0.029	0.037	0.355	0.070	0.000	0.000	0.337	0.143	0.000	0.000
	V10	V11	V12	V13	V14	V10	V11	V12	V13	V14
Min.	-0.002	-0.038	0.000	0.000	0.000	0.000	0	0.000	0.024	0.000
1^{st} Qu.	0.000	0.000	0.030	0.110	0.000	0.000	0.000	0.143	0.080	0.000
median	0.000	0.000	0.094	0.125	0.000	0.000	0.000	0.199	0.090	0.000
mean	0.001	0.003	0.088	0.113	0.005	0.001	0.000	0.176	0.083	0.001
3^{rd} Qu.	0.000	0.000	0.136	0.135	0.000	0.000	0.000	0.245	0.101	0.000
Max.	0.089	0.062	0.246	0.152	0.089	0.116	0.101	0.344	0.110	0.029
	V15	V16	V17	V18	V19	V15	V16	V17	V18	V19
Min.	0.000	0.000	-0.003	-0.052	-0.119	0.000	0.000	-0.068	-0.182	-0.136
1^{st} Qu.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
median	0.027	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
mean	0.039	0.012	0.028	0.010	-0.003	0.006	0.006	0.024	0.000	-0.001
3^{rd} Qu.	0.058	0.000	0.032	0.000	0.000	0.000	0.000	0.045	0.000	0.000
Max.	0.602	0.489	0.221	0.222	0.000	0.520	0.440	0.147	0.113	0.000

$$y^*(TCC_{Bengaluru}) = 0.213 - 0.0034V6 + 0.029V12^* + 0.23V13^* - 0.65V14^* + 0.063V15^* + 0.031V16^*$$

$$y^*(Deaths_{Bengaluru}) = 1.84 + 0.004V1^* + 0.0076V8^* + 0.043V9^* + 0.0027V13^* + 0.032V16^* + 0.0038V17^*$$

6. Discussion

The current research implemented spatial analysis techniques to analyse 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 of 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 same 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. In response to this concern, the ridge and LASSO techniques were initially employed on the spatial models. It was discovered that among the spatial models implemented in this research, the GWL model exhibited superior performance. By integrating spatially varying coefficients, the GWL model effectively captured localized fluctuations and heterogeneity in the association between the dependent and independent variables, while also addressing collinearity concerns among the independent variables. Although the GWL model generates different models and identifies significant independent variables for different locations, this study also determined the independent variables that, on average, influence COVID-19 cases and deaths across Indian districts.

The findings of the GWL model demonstrated a positive relationship between the 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 (Chen *et al.* (2020), Tosepu *et al.* (2020), Bashir *et al.* (2020)). 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 and deaths, 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. 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. Consistent with prior investigations, the proportion of the population residing in urban settings and the density of specific districts exerted anticipated effects on COVID-19 incidence and mortality within those particular areas. Conversely, certain variables like relative humidity, household access to electricity, and possession of health insurance exhibited negligible influence on COVID-19 patterns in Indian districts. Furthermore, no specific age group demonstrated disproportionate susceptibility to COVID-19; however, elderly individuals were identified as having an elevated risk of mortality attributed to the virus.

The GWL model introduces spatial variability in coefficients, capturing differences in various locations. The range of coefficients gives insight into how relationships between vari-

ables change across space. The GWL model's findings help understand the degree to which the identified risk factors account for differences in COVID-19 cases and deaths in diverse districts. For instance, districts like Mumbai, Chennai, Pune, Kolkata, Sagar, Jabalpur, Narshimpur, Raisen, Porbandar, Junagarh, and Somnath exhibit significant variation in COVID-19 outcomes (ranging from 80 to 86 percent). This highlights the strong impact of the identified risk factors in these areas. Conversely, the considered variables struggle to explain variations in certain districts, particularly in parts of Punjab (such as Bhatinda, Faridkot, Moga) and the northeastern region. Similarly, regions like Sirsa, Panchkula, and districts in Himachal Pradesh, JK, and Ladhak have limited explanatory capability. 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.

7. 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 analyse 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 districts in India. Additionally, this study addressed the issue of multicollinearity in spatial models through the utilization of ridge and LASSO techniques. 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. Significantly, the study brought to light the influence of distinct district factors on the occurrence of COVID-19. These factors encompass sanitation facilities, accessibility to healthcare, pre-existing medical conditions like high blood pressure and diabetes, women's educational levels, rates of tobacco and alcohol consumption, climatic conditions, and the presence of undernourished women. Moreover, the research established that older populations are at a heightened risk of mortality following infection with COVID-19. 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.

Acknowledgements

We would like to express my sincere gratitude to Banasthali Vidyapith for providing the necessary resources and support throughout the duration of this research. The academic environment and facilities at the university have played a crucial role in the successful completion of this study.

References

Adekunle, I. A., Onanuga, A. T., Akinola, O. O., and Ogunbanjo, O. W. (2020). Modelling spatial variations of coronavirus disease (COVID-19) in Africa. *Science of the Total Environment*, **729**, 138998.

Ahmed, F., Ahmed, N., Pissarides, C., and Stiglitz, J. (2020). Why inequality could spread Covid-19. *The Lancet Public Health*, **5**, e240.

Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis*, **27**, 93–115.

Appiah-Otoo, I. and Kursah, M. B. (2022). Modelling spatial variations of novel coronavirus disease (Covid-19): evidence from a global perspective. *GeoJournal*, **87**, 3203–3217.

Bashir, M. F., Ma, B., Komal, B., Bashir, M. A., Tan, D., Bashir, M., *et al.* (2020). Correlation between climate indicators and Covid-19 pandemic in New York, USA. *Science of the Total Environment*, **728**, 138835.

Baye, M. R. and Parker, D. F. (1984). Combining ridge and principal component regression: a money demand illustration. *Communications in Statistics-Theory and Methods*, **13**, 197–205.

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

Chandra, S. and Sarkar, N. (2016). A restricted r-k class estimator in the mixed regression model with autocorrelated disturbances. *Statistical Papers*, **57**, 429–449.

Chen, B., Liang, H., Yuan, X., Hu, Y., Xu, M., Zhao, Y., Zhang, B., Tian, F., and Zhu, X. (2020). Roles of meteorological conditions in Covid-19 transmission on a worldwide scale. *MedRxiv*, **3**, 2020–03.

Chung, H. W., Apio, C., Goo, T., Heo, G., Han, K., Kim, T., Kim, H., Ko, Y., Lee, D., Lim, J., *et al.* (2021). Effects of government policies on the spread of Covid-19 worldwide. *Scientific Reports*, **11**, 20495.

Dutta, I., Basu, T., and Das, A. (2021). Spatial analysis of Covid-19 incidence and its determinants using spatial modeling: A study on India. *Environmental Challenges*, **4**, 100096.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Kejian, L. (1993). A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods*, **22**, 393–402.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, **382**, 1199–1207.

Pereira, M. and Oliveira, A. M. (2020). Poverty and food insecurity may increase as the threat of Covid-19 spreads. *Public Health Nutrition*, **23**, 3236–3240.

Ramírez-Aldana, R., Gomez-Verjan, J. C., and Bello-Chavolla, O. Y. (2020). Spatial analysis of covid-19 spread in Iran: Insights into geographical and structural transmission determinants at a province level. *PLoS Neglected Tropical Diseases*, **14**, e0008875.

Rocklöv, J. and Sjödin, H. (2020). High population densities catalyse the spread of Covid-19. *Journal of Travel Medicine*, **27**, taaa038.

Sardar, S., Abdul-Khaliq, I., Ingar, A., Amaidia, H., and Mansour, N. (2020). Covid-19 lockdown: A protective measure or exacerbator of health inequalities? a comparison between the United Kingdom and India. a commentary on "the socio-economic implications of the coronavirus and covid-19 pandemic: A review. *International Journal of Surgery (London, England)*, **83**, 189.

Sarkar, S. K., Ekram, K. M. M., and Das, P. C. (2021). Spatial modeling of Covid-19 transmission in Bangladesh. *Spatial Information Research*, **29**, 715–726.

Sridhar, K. S. (2023). Urbanization and Covid-19 prevalence in India. *Regional Science Policy & Practice*, **15**, 493–505.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**, 267–288.

Tosepu, R., Gunawan, J., Effendy, D. S., Lestari, H., Bahar, H., Asfian, P., *et al.* (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Science of the Total Environment*, **725**, 138436.

Wheeler, D. and Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, **7**, 161–187.

Wheeler, D. C. (2007). Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and Planning A*, **39**, 2464–2481.

Wheeler, D. C. (2009). Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. *Environment and Planning A*, **41**, 722–742.