Does the Indo-Gangetic Plain, characterized by high pollution levels and poor public healthcare infrastructure, exhibit a clear pattern of increased COVID-19 impact compared to the rest of India?

Kishika Mahajan

March 2025

Abstract

The COVID-19 pandemic has exposed vulnerabilities in healthcare systems worldwide, with regions characterized by high pollution and poor healthcare infrastructure facing unique challenges. This study examines the spatial patterns of COVID-19 mortality in India, focusing on the Indo-Gangetic Plain, a region known for its severe air pollution and limited healthcare resources. Using district-level data, we explore the relationship between air pollution, healthcare infrastructure, age distribution, and migration trends to understand regional disparities in COVID-19 impact. We employ Principal Component Analysis (PCA) and K-Medoids clustering to identify patterns and clusters of COVID-19 mortality. Our findings reveal that age is the most significant determinant of COVID-19 mortality, overshadowing the effects of pollution and healthcare infrastructure. Contrary to expectations, the Indo-Gangetic Plain does not exhibit the highest mortality rates, despite its unfavorable conditions. Instead, regions with older populations and higher migration activity experience the most severe outcomes. These results highlight the importance of considering demographic factors in public health planning and provide actionable insights for reducing COVID-19 mortality in future health crises. The study underscores the need for targeted interventions for older populations, improved healthcare infrastructure, and better management of migration patterns to mitigate the impact of future pandemics.

1 Introduction

The COVID-19 pandemic has had a profound impact on global health, economies, and societies, exposing vulnerabilities in healthcare systems and highlighting the role of environmental and demographic factors in shaping health outcomes. India, with its vast population and diverse regional characteristics, has been one of the most affected countries. Within India, the Indo-Gangetic Plain, stretching across northern and eastern states such as Punjab, Haryana, Delhi, Uttar Pradesh, Bihar, and West Bengal, is home to some of the most polluted cities in the world, with particulate matter (PM2.5) levels consistently exceeding safe limits.

The state of the health infrastructure of countries was also revealed during the pandemic. The Indo-Gangetic Plain was also seen to have poor public health infrastructure which would further worsen the impact of the pandemic.

This study seeks to address the interplay between a multitude of factors by examining the spatial patterns of COVID-19 mortality in India, with a particular focus on the Indo-Gangetic Plain. Using district-level data, the aim is explore the relationship between air pollution, healthcare infrastructure and migration trends to understand why certain regions experienced higher COVID-19 mortality rates than others.

Specifically, the study aims to answer the following question: Does the Indo-Gangetic Plain, characterized by high pollution levels and poor public healthcare infrastructure, exhibit a clear pattern of increased COVID-19 impact compared to the rest of India? The hypothesis is that long-term exposure to high levels of air pollution contributes to chronic respiratory conditions and weakened immune responses, thereby increasing the severity and mortality of COVID-19. Consequently, areas with higher pollution levels and lower healthcare infrastructure, such as the Indo-Gangetic Plain, are expected to experience more intense COVID-19 mortality rates compared to areas with lower pollution levels and better healthcare infrastructure.

To test this hypothesis, the study employs spatial analysis techniques, including Principal Component Analysis (PCA) and K-Medoids clustering, to identify patterns and clusters of COVID-19 mortality across India. By integrating multiple datasets—ranging from air pollution levels and healthcare infrastructure to age distribution and migration trends—the study provides a comprehensive understanding of the factors driving regional disparities in COVID-19 impact. This analysis not only sheds light on the unique challenges faced by the Indo-Gangetic Plain but also offers broader insights into the interplay between environmental, healthcare, and demographic factors in shaping health outcomes during a pandemic.

In the following sections, the data and methods used in the analysis are described in detail, the findings are presented, and the implications for public health policy and future research are discussed. This study aims to deepen the understanding of the factors that shaped India's COVID-19 experience and provide actionable insights for building more resilient health systems in the face of future challenges.

2 Data Description and Processing

This study utilizes six datasets[1] to analyze the impact of COVID-19 in India. Five of these datasets, including shapefiles, are merged to create a master dataset for clustering and spatial analysis. The integration is performed using a unique district ID assigned to each district in India. After reviewing all the datasets, complete data is available for 531 districts. Jammu and Kashmir, Gujarat, and Lakshadweep are entirely excluded from the analysis due to missing data for certain atrributes in these states. The projection used to project the data for the purposes of clustering was WGS 84 (EPSG:4326).

1. **Shapefiles**: A district-level shapefile of India is used to facilitate spatial analysis at the district level.
2. **Average PM2.5**: The study employs district-wise average PM2.5 levels as a proxy for pollution. The data is available from 1998 to 2020. Since the first case of Covid was registered in January, 2020 in India, the average PM2.5 levels considered for the purposes of this study are from 2017 to 2019.

[1] All datasets were obtained from the Development Data Lab.

3. **Hospitals**: The hospital dataset provides four key indicators that collectively reflect the healthcare capacity in each district[2] in each district. These are as follows:
   (a) Total number of beds
   (b) Total number of beds with oxygen cylinders
   (c) Total staff in healthcare facilities
   (d) Total public healthcare facilities
4. **Age Distribution**: The dataset includes population shares for different age groups like share of total population in the age group of 0-4 years, Share of total population in the age group of 5-9 years and so on till Share of total population in the age group of 80-84 years. To obtain a representative measure, a weighted average age is computed for each district.
5. **Migration Data**: The migration dataset categorizes migration patterns as follows:
   (a) Long-term inside the district migration
   (b) Short-term inside the district migration
   (c) Long-term outside the district migration
   (d) Short-term outside the district migration

   The dataset reports the share of migrants in each category relative to the national migration figures.
6. **Covid Deaths**: This dataset tracks COVID-19 deaths across India from January 30, 2020, to October 31, 2021, and serves as the primary measure of COVID-19 impact in this study[3].

3 Methodology

3.1 Exploratory Data Analysis

Before moving onto clustering and analyses, we conduct exploratory data analyses to establish spatial auto-correlations and hence uncover the the spatial patterns of the data. To motivate out hypothesis, we conduct a univariate analyses wherein the aim is to focus on one variable at a time (average PM2.5 and each component of health infrastructure.)

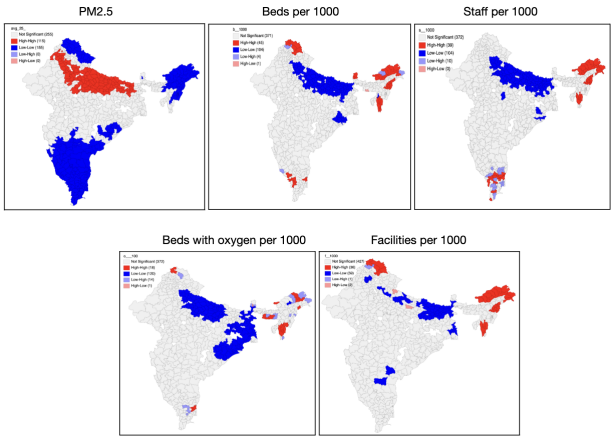

Figure 1: Univariate Local Moran's I analysis

As observed from the Local Moran's I analysis, the Indo-Gangetic Plain exhibits high-high clusters for average PM2.5 levels, indicating that the region has extremely high pollution levels. This spatial clustering

[2]For the purposes of this study, we obtain the per 1000 measure of each of these variables.
[3]For the purposes of this study, we obtain the per 1000 measure of the deaths.

aligns with existing knowledge about the severe air pollution in this area, which is home to some of the most polluted cities in the world. Furthermore, the analysis reveals low-low clusters for all healthcare infrastructure indicators in the Indo-Gangetic Plain, highlighting the region's poor public health infrastructure. These findings underscore the challenges faced by the region in terms of healthcare access and capacity, which may exacerbate the impact of health crises such as the COVID-19 pandemic.

The next part of this section focuses on clustering the regions based on the attributes discussed so far. The two main methods used for this analysis are Principal Component Analysis (PCA) and K-Medoids.

3.2 Principal Component Analysis

Principal Component Analysis will be used to reduce the dimensionality of the health infrastructure and migration variables. PCA is chosen as a method as the data has multiple components and doing this can retain the variation in the data for health infrastructure and migration in a much simpler and interpretable manner. Further, by reducing the dimensionality of the data, PCA will ensure that the clustering is not overwhelmed by too many variables, making it easier to identify meaningful patterns. We run two sets of PCA, one for the four variables that together comprise the level of health infrastructure, and the second on the four variables that together comprise the migration trends.

PCA on health infrastructure indicators: After running the PCA on the four health infrastructure indicators, we get the following variable loadings:

	PC1	PC2	PC3	PC4
b_1000	0.562846	-0.0540383	-0.253926	-0.784733
f_1000	0.522898	0.289873	-0.587608	0.545224
s_1000	0.474009	0.489332	0.728628	0.0705122
o_100	0.430231	-0.820737	0.243599	0.286274

As per the Kaiser criterion, we retain just the first principal component which explains 63% of the total variance. As can be understood from the loadings of the first principal component (PC1) in the table, a higher value of PC1 would imply higher quality in all health infrastructure measures.

PCA on migration indicators: After running the PCA on the four migration indicators, we get the following variable loadings:

	PC1	PC2	PC3	PC4
inltmgr	0.632668	0.233072	0.105998	0.730871
instmgr	0.610059	0.197127	0.407845	-0.650101
otltmgr	0.46633	-0.386577	-0.777824	-0.167585
otstmgr	0.100479	-0.870274	0.46628	0.122925

As per the Kaiser criterion, we retain the first two principal components which cumulatively explain 78.5% of the total variance. As can be understood from the loadings of the first principal component (PC1) in the table, PC1 shows overall migration trends which means that a higher PC1 value shows that the overall migration in a district is high, both long- and short-term, as well as both inside and outside the district.
As for the second principal component (PC2), it shows a comparison of in and out of the district migration. More particularly, a higher value of PC2 implies a high level of in-migration while a low level of PC2 suggests more out-migration from the district.

3.3 K-Medoids Clustering

After performing the principal component analyses, the districts were clustered using K-Medoids on the basis of the retained principal components, average PM2.5 measures and the centroids.

This will allow us to identify spatial patterns and assess whether districts in the Indo-Gangetic Plain form distinct clusters with higher COVID-19 mortality rates. K-medoids will choose one of the actual districts as cluster centers with the chosen district being representative of the cluster increasing the interpretability significantly. K-Medoids is also less sensitive to outliers. This is particularly important given the presence of extreme values in variables like pollution levels and migration rates. Lastly, to ensure the clustering is also contiguous, relative weights will be assigned to both coordinates and attributes.

3.3.1 Selection of spatial weights for clustering

We can first look at clusters based on the aforementioned variables without including the coordinates as a metric of clustering. We can first look at the clustering without assigning weights to the coordinates.

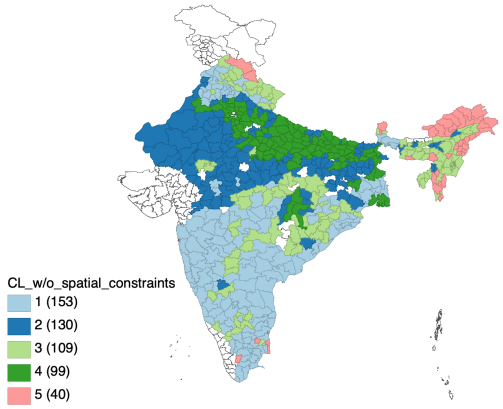

Figure 2: Clustering without including spatial components

As can be seen from the above figure, the clusters are not spatially constrained. To enforce spatial constraints in clustering, different weights were assigned to the coordinates. After experimentation, a weight of 0.45 was assigned to the coordinates for clustering.

4 Results and Analysis

4.1 Initial Clustering

4.1.1 Selection of number of clusters

We cluster using the average PM2.5, the retained principal components and the coordinates (assigning 0.45 as the weight for coordinates). To choose the number of clusters, we use the Elbow method as shown below:

Number of Clusters (k)	Total Within-Cluster Sum of Squares (WCSS)
3	5083.42
4	4061.76
5	3843.09
6	3721.42
7	3640.54

Table 1: Total Within-Cluster Sum of Squares

As can be seen, the *elbow* is at 5 clusters as after that the fall in the WCSS is minimal. Hence, we choose to retain 5 clusters for the analysis. After forming 5 clusters using the average PM2.5, the retained principal components and the coordinates (while assigning 0.45 as the weight for coordinates), we obtain the following clusters:

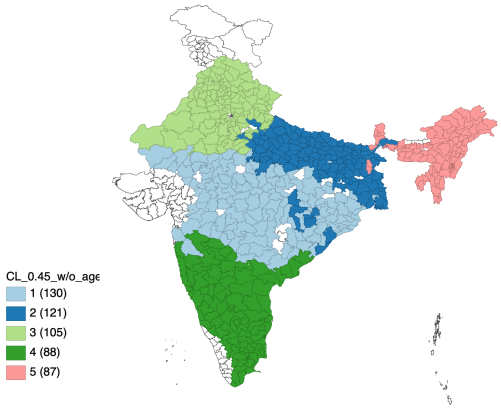

Figure 3: Clustering on the basis of pollution and the retained principal components.

While not entirely covered, the Indo-Gangetic plain majorly falls in Cluster 2 and some parts of Cluster 3. So, we can begin by looking at the characteristics of each of these clusters.

Table 2: Clusters' Description

Cluster	**PM2.5**	**Health Infra**	**Overall Migration**	**In/Out Migration**	**Interpretation**
C1	47.39	-0.33	-0.40	-0.55	• Moderate air pollution • Poor health infrastructure • Low migration activity • More people leaving than coming
C2	97.56	-0.97	-0.13	-0.61	• Highest air pollution • Worst health infrastructure • Low migration activity overall • Significant in-migration

Continued on next page

Table 2 – *Continued from previous page*

Cluster	PM2.5	Health Infra	Overall Migration	In/Out Migration	Interpretation
C3	73.94	-0.48	-0.03	0.77	• High air pollution • Poor health infrastructure • Lower migration activity • More in-migration than out-migration
C4	31.29	0.08	0.54	0.03	• Lowest air pollution • Best health infrastructure • Highest migration activity • Balanced in-migration and out-migration
C5	31.64	0.06	-0.92	0.61	• Low air pollution • Good health infrastructure • Very low migration overall • More in-migration than out-migration

Upon a very basic examination, we can see that ideally, cluster 4 seems to be the most favorable cluster to be in with the best health infrastructure and the lowest levels of pollution. Hence, we can assume that this cluster would have experienced the least amount of deaths in the pandemic. So, we look at the mean deaths in each cluster and obtain the following results:

Cluster	Mean Deaths per 1000
1	17.459972
2	7.079592
3	20.197495
4	34.813943
5	4.247341

Table 3: Mean Deaths per Cluster

As can be seen, the results are contrary to what we would have expected from our cluster descriptions.

- Firstly, we would have expected Cluster 4 (the southern part of India) to have the least number of mean deaths, however it seems to be having the most number of mean deaths.
- Secondly, the Indo-Gangetic Plain characterised mostly by Cluster 2 and somewhat by Cluster 3, which is characterised by the worst pollution levels and the worst health infrastructure, actually observes way lesser deaths as compared to Cluster 4.

This reinforces the idea that pollution levels and public health infrastructure alone cannot determine the impact of Covid-19 deaths. While overall migration can be seen to be the highest in Cluster 4, we need to account for different proportions of in and out migrations as well. Hence, it can be concluded that there is some other factor that influenced the impact of Covid-19 in India.

4.2 Clustering with Age

Several studies have shown that Covid was disproportionately harmful to the elderly, with the elderly being more vulnerable to the physical effects of COVID-19 and significantly more likely to succumb to COVID-19.

Hence, we include the average age of each district as a metric in our clustering along with all the previous variables used in the clustering. Once again, we continue to retain the 0.45 weight for the coordinates.

We use the elbow method to choose the number of clusters after including Age and all the previous attributes.

Number of Clusters (k)	**Total Within-Cluster Sum of Squares (WCSS)**
3	5325.48
4	4742.40
5	4289.77
6	4162.06
7	4045.79

Table 4: Total Within-Cluster Sum of Squares for clusters including Age

As can be seen, the elbow is at 5 clusters as after that the fall in the WCSS is minimal. Hence, we choose to retain 5 clusters for the analysis. After forming 5 clusters using the average age, average PM2.5, the retained principal components and the coordinates (while assigning 0.45 as the weight for coordinates), we obtain the following clusters:

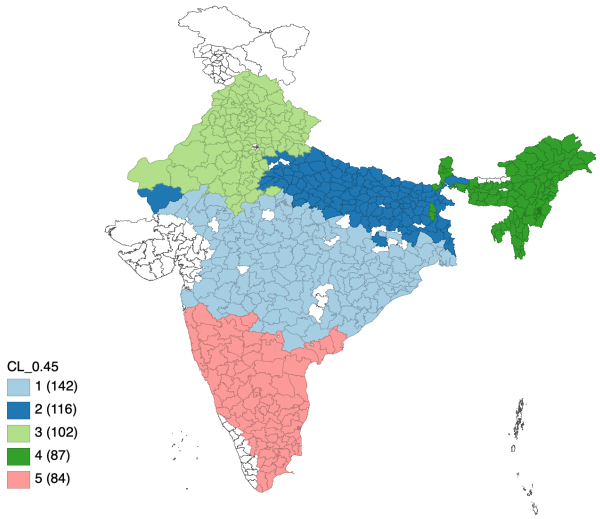

Figure 4: Clustering after including age distributions

The first important thing to note here is that in this case, Cluster 2 is more representative of the Indo-Gangetic plain. Once again, we can look at the descriptions of the clusters obtained.

Table 5: Clusters' Description

Cluster	Average Age	Average PM2.5	Health Infrastructure	Overall Migration	In/Out Migration	Interpretation
C1	29.31	42.88	-0.347	-0.203	0.065	• Young population • Medium PM2.5 • Bad health infrastructure • Low overall migration
C2	27.3	100.29	-0.947	-0.351	-0.423	• Young population • Highest PM2.5 • Worst health infrastructure • Low overall migration • Most out-migration, least in-migration
C3	30.17	73.94	-0.478	-0.032	0.769	• Medium age population • High PM2.5 levels • Bad health infrastructure • Moderate overall migration • Most in-migration, least out-migration
C4	25.92	31.64	0.065	-0.924	0.608	• Youngest population • Low PM2.5 • Good health infrastructure • Lowest overall migration activity • High in-migration, relatively low out-migration
C5	32.56	31.29	0.08	0.538	0.034	• Oldest population • Lowest PM2.5 • Best health infrastructure • Highest overall migration activity

We can make two primary conclusions:

1. The Indo-Gangetic Plain that is showcased by Cluster 2 can be seen to have the worst pollution level and the worst public health infrastructure level. It should also be noted that this cluster has the second youngest population of all the clusters and the most **out** migration.

2. In contrast, Cluster 5 which shows the southern part of India has the best public health infrastructure

and the lowest pollution levels. However, this cluster is also characterised by the oldest population of all the other clusters and the highest overall migration.

We can now look at the mean deaths per 1000 in each cluster to assess if age explains the results previously observed.

Cluster	Mean Deaths per 1000
1	16.912421
2	5.873190
3	20.430685
4	4.247341
5	37.490308

Table 6: Mean Deaths per Cluster

As can be seen from the above table, once again, Cluster 2 has extremely low death rates while on the other hand, Cluster 5 has the highest death rates. Hence, our hypothesis that age played a significantly important role in the differential impact of Covid-19 is indeed true.

5 Discussion

The findings of this study provide valuable insights into the factors driving regional disparities in COVID-19 mortality across India, with a particular focus on the Indo-Gangetic Plain. Contrary to our initial hypothesis, the Indo-Gangetic Plain—characterized by high pollution levels and poor healthcare infrastructure—does not exhibit the highest COVID-19 mortality rates. Instead, regions with older populations and higher migration activity, experience the most severe outcomes. These results underscore the complex interplay between environmental, healthcare, and demographic factors in shaping COVID-19 impact.

5.1 Key Findings

1. **Role of Age in COVID-19 Mortality:** The most striking finding is the significant role of age in determining COVID-19 mortality. Cluster 5, which has the oldest population, records the highest mortality rates despite having low pollution levels and the best healthcare infrastructure. This aligns with global evidence that older populations are more vulnerable to severe COVID-19 outcomes due to weaker immune responses and higher prevalence of comorbidities.
   In contrast, the Indo-Gangetic Plain (Cluster 2), with the second youngest population, experiences lower mortality rates despite its high pollution levels and poor healthcare infrastructure. This suggests that demographic factors, particularly age distribution, may outweigh the effects of pollution and healthcare access in determining COVID-19 outcomes.

2. **Role of Migration in COVID-19 Mortality:** The role of migration in COVID-19 mortality is another critical factor highlighted by the analysis. Cluster 5, which has the highest overall migration activity, likely experienced an increased spread of the virus due to higher population mobility. This could explain its elevated mortality rates despite having low pollution levels and the best healthcare infrastructure. Further, although C4 has a lower average age than C2, the deaths are not that very different. This can be attributed to the in and out migration differences. C2 observes high out migration than in migration which means it may have benefited from reduced population density during the pandemic, contrary to C4 which observes higher in more in migration.

3. **Pollution and Healthcare Infrastructure:** While pollution and healthcare infrastructure are important determinants of health outcomes, their impact on COVID-19 mortality appears to be secondary to age in this study. High pollution levels in the Indo-Gangetic Plain did not translate into higher mortality rates, possibly because the region's younger population was less susceptible to severe COVID-19. However, this does not diminish the importance of addressing pollution and healthcare disparities.

Long-term exposure to high pollution levels can exacerbate chronic respiratory and cardiovascular conditions, while poor healthcare infrastructure limits a region's capacity to respond to health crises.

5.2 Policy Implications

1. **Targeted Interventions for Older Populations:**
   The study highlights that age is the most significant determinant of COVID-19 mortality, with older populations experiencing the most severe outcomes. This finding underscores the need for targeted interventions to protect vulnerable demographic groups. This should particularly include strengthening healthcare infrastructure in regions with older populations, including the expansion of hospital beds, oxygen supplies, and critical care facilities.

2. **Addressing Pollution and Healthcare Disparities:** While pollution and healthcare infrastructure were not the primary drivers of COVID-19 mortality in this study, addressing these issues remains critical for long-term public health. Stricter pollution controls should be implemented as the people residing in high population regions are not just more susceptible to threats during the time of the pandemic, everyday, the life expectancy is reduced due to the long lasting effects of pollution.
   It is also pertinent to note the varying levels of public health infrastructure in the country. Perhaps, it is only because the Southern India had high quality public infrastructure that the deaths were what they were seen; the impact would've been extreme if the Indo-Gangetic plain had the oldest population.

3. **Migration and Urban Planning:** Migration trends were found to influence COVID-19 outcomes, with regions experiencing high in-migration rates facing higher mortality rates. Policies to manage migration flows and reduce the strain on healthcare systems in destination regions should be focused upon. For example, incentivizing economic development in source regions can reduce out-migration and promote balanced regional growth.

5.3 Limitations and Future Research

It is important to acknowledge the limitations of this study, which can inform future research directions and improve the robustness of spatial analyses of COVID-19 outcomes.

1. **Reporting Rates:**
   While the study identifies higher COVID-19 mortality rates in southern India compared to other regions, this observation may be influenced by disparities in reporting. Southern India, with its relatively better state capacity, may have more accurate and comprehensive reporting of COVID-19 deaths. In contrast, underreporting in other regions, such as the Indo-Gangetic Plain, could obscure the true impact of the pandemic. Future studies should consider incorporating data on reporting rates to uncover potential spatial patterns and provide a more accurate picture of COVID-19 mortality.

2. **Vaccination and Long-Term Effects:**
   The study does not account for vaccination rates or post-pandemic recovery, which could significantly influence COVID-19 outcomes. For instance, regions with higher vaccination coverage may have experienced lower mortality rates, while areas with better healthcare infrastructure may have facilitated faster recovery. Additionally, the long-term effects of high pollution levels in the Indo-Gangetic Plain could make its population more susceptible to future health threats, even if the immediate impact of COVID-19 was mitigated by demographic factors such as a younger population. Future research should explore these dimensions to provide a more comprehensive understanding of the pandemic's impact.

3. **Private Healthcare Infrastructure:**
   There is also the possibility that the state of healthcare infrastructure in the Indo-Gangetic Plain is not as poor as the data suggests, due to the presence of significant private healthcare infrastructure. While this is a plausible consideration, it is equally important to note that the southern regions of India also have a high level of private healthcare infrastructure. Including data on private healthcare facilities in future analyses could provide a more comprehensive picture of healthcare capacity and potentially reveal clearer patterns in COVID-19 outcomes. This would help disentangle the relative contributions of public and private healthcare systems to pandemic response and mortality rates.

6 Conclusion

This study focuses on understanding the factors driving regional disparities in COVID-19 mortality across India, particularly in the Indo-Gangetic Plain. By integrating spatial analysis techniques such as PCA and K-Medoids clustering, it was identified that age distribution and migration patterns play a more significant role in shaping COVID-19 outcomes than pollution or healthcare infrastructure. Contrary to the initial hypothesis, the Indo-Gangetic Plain, despite its high pollution levels and poor public healthcare infrastructure, does not exhibit the highest mortality rates. Instead, regions with older populations and higher in-migration, such as Cluster 5, experience the most severe outcomes. These findings challenge conventional assumptions about the primary drivers of pandemic mortality and highlight the complex interplay between environmental, healthcare, and demographic factors.

The implications of this research extend beyond COVID-19, offering valuable lessons for public health policy and pandemic preparedness. Policymakers must prioritize targeted interventions for older populations, invest in healthcare infrastructure, and address the long-term effects of pollution. Additionally, urban planning strategies that account for migration patterns and population density can help mitigate the spread of infectious diseases in high-risk regions. By leveraging data-driven approaches and spatial analysis, this study not only provides actionable insights for addressing COVID-19 but also lays the groundwork for building more resilient health systems capable of responding to future health crises. Ultimately, this research underscores the importance of a holistic approach to public health, one that considers the multifaceted factors shaping health outcomes in diverse regional contexts.